Predict Clicked Ads Customer Classification by Using Machine Learning

Created by:
Raffel Ravionaldo
raffelrazor@gmail.com
https://www.linkedin.com/in/raffel-ravionaldo/

A fresh graduate interested in data, learns about data through 2 data science bootcamp, first organized by Rakamin and the second by binar academy. To deepen my knowledge at data, I took part in a virtual intership by rakamin cooperating with ID/X Partner, Home Credit Indonesia and kimia farma, and a virtual project organized by forage.com cooperating with British Airways.

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Overview

"A company in Indonesia wants to know the effectiveness of an advertisement that they display, this is important for the company to be able to know ho much the advertising has been achieved so that it can attract customers to see the advertisement.

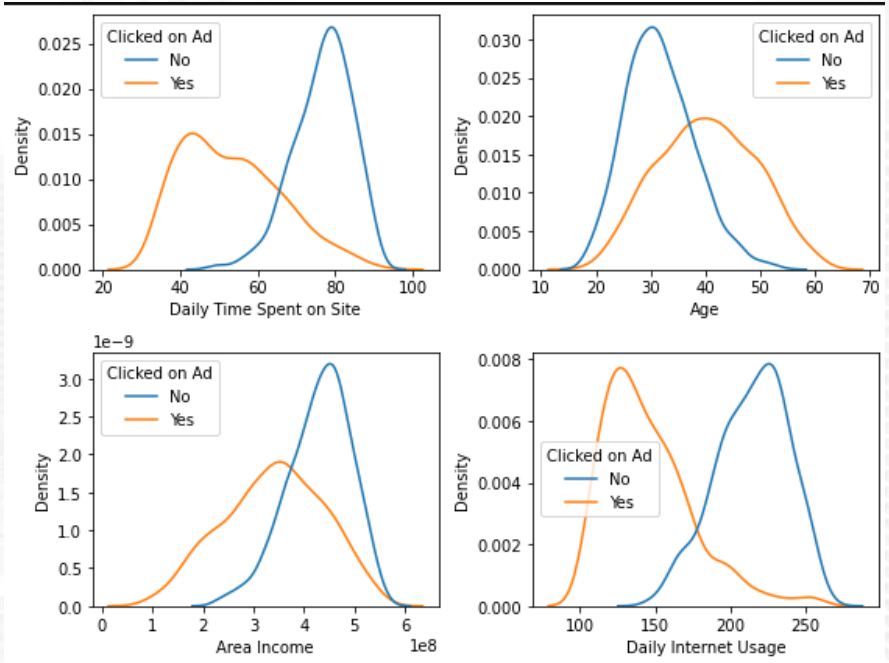
By processing historical advertising data and finding insights and patterns that occur, it can help companies determine marketing targets, the focus of this case is to create a machine learning classification model that functions to determine the right target customers."

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Rakamin Academy

Customer Type and Behaviour Analysis on Advertisement

Univariate Analysis

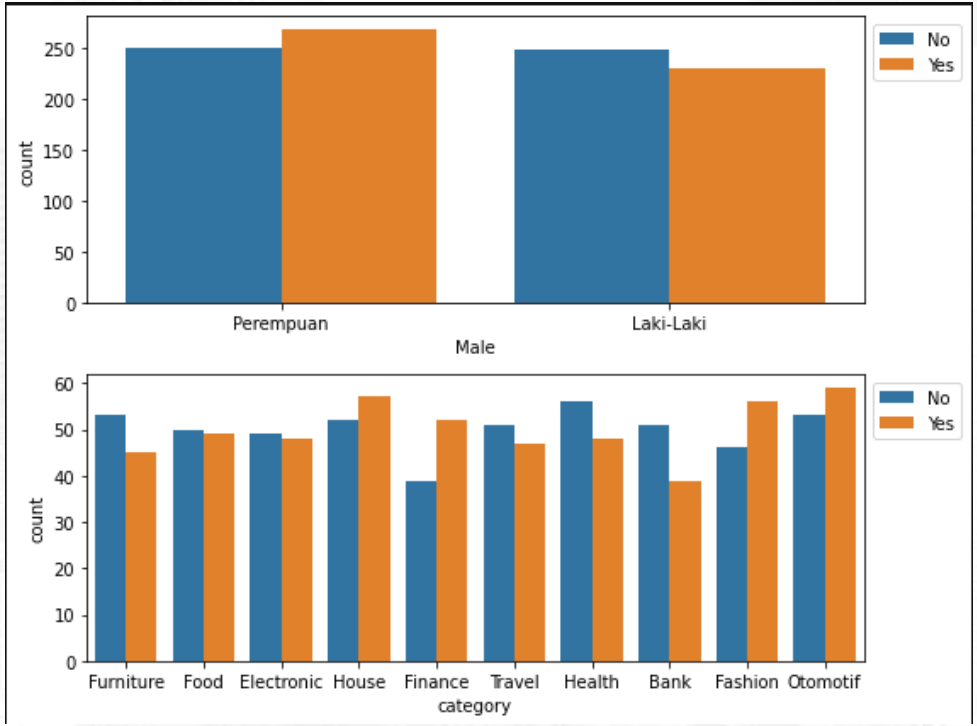


From the graph, we know that customer tend to click our ad if they are :

1. Have <= 60 daily time spent on our site.
2. Are >=40 years old.
3. Have an income <= 3.2 hundred million
4. And have <= 170 daily internet usage.

Want to see the code? Click Here

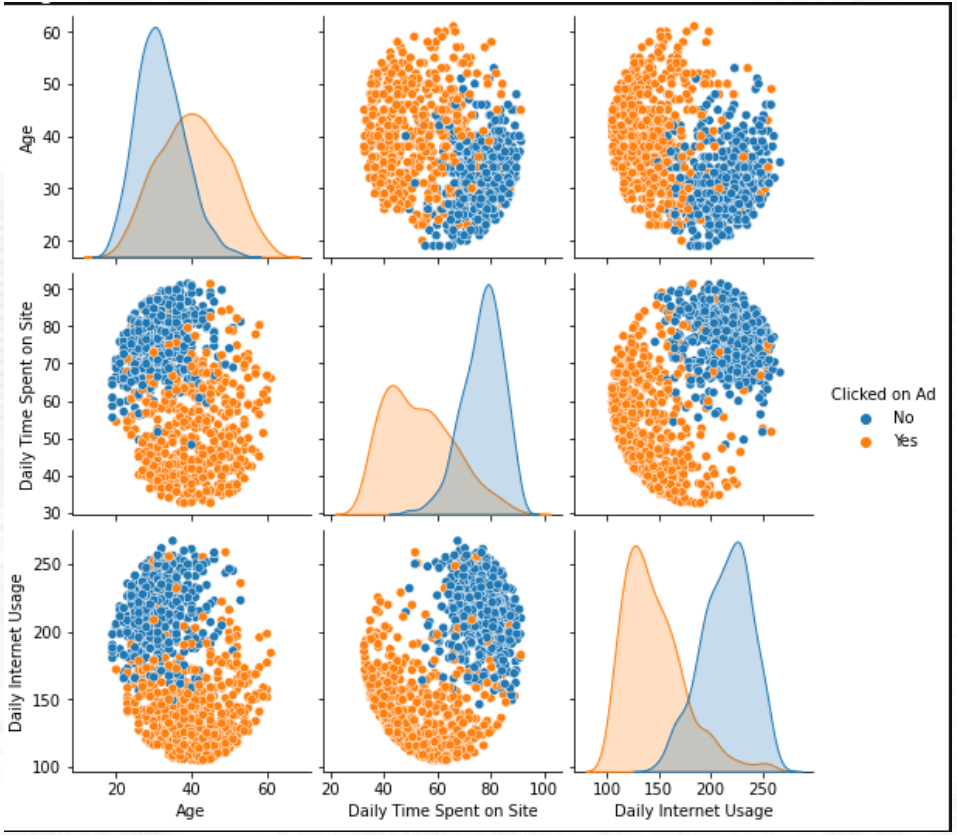
Univariate Analysis



From the graph, we know that woman have a percentage above 50% to click on our ads, and customer is more interested with our ads about house, finance, fashion and automotive,this can be said because of the percentage customer click our ads above 50%.

Want to see the code? Click Here

Bivariate Analysis

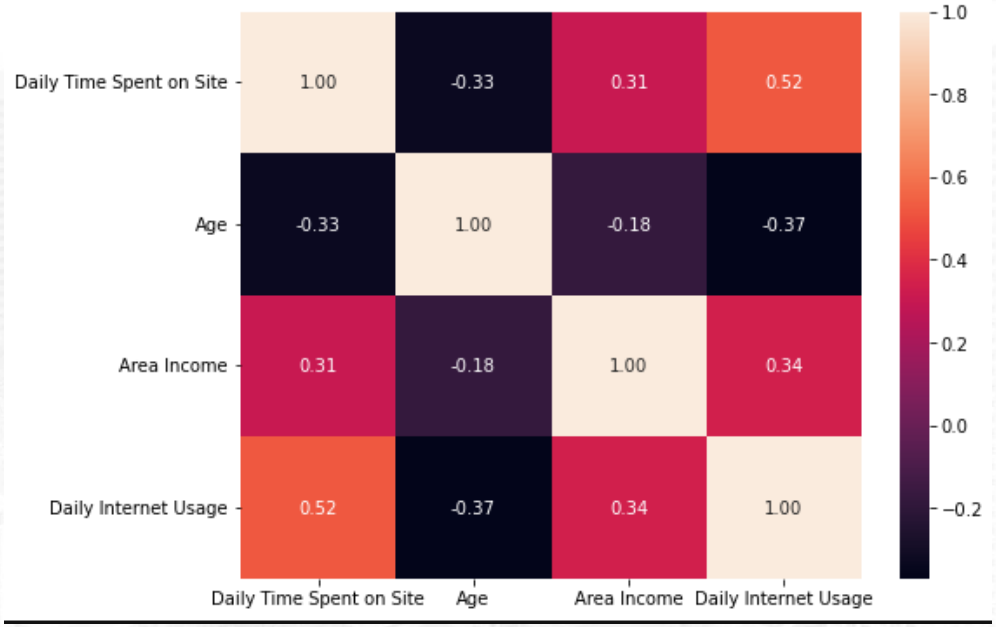


From the graph, we know :

1. The older our customer, have a few daily time spent on site and daily internet usage they tend to click our ads.
2. The younger our customer, have a more daily time spent on site and daily internet usage they tend to not click our ads.

Want to see the code? Click Here

Multivariate Analysis



From graph, we know that Daily Internet Usage and Daily Time spent on site have a highest correlation between another column, the second highest correlation is Age with Daily Internet Usage and the third is area income with daily internet usage.

Want to see the code? Click Here

Multivariate Analysis

	Column	Category column	P_value from chi square
10	city	province	0.000000
5	Clicked on Ad	city	0.206285
3	Male	category	0.232151
0	Male	Clicked on Ad	0.296170
11	city	category	0.362574
13	province	Clicked on Ad	0.380631
6	Clicked on Ad	province	0.380631
2	Male	province	0.430435
1	Male	city	0.467334
8	city	Male	0.467334
15	province	category	0.607245
7	Clicked on Ad	category	0.695155
17	category	Clicked on Ad	0.695155

Using Chi-Square for know association from category column, we know that clicked on Ad have a high correlation with city and Male with category.

Want to see the code? Click Here

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Rakamin Academy

Data Cleaning & Preprocessing

Data Preprocessing

Rakamin Academy

Missing Value

4 columns have null data,
3 of them is numeric so we fill it with median of that data and
Other we fill it mode
We don't have duplicate data

Feature Encoding

Extract Datetime Data

We get 5 new column, there are year, month, day, weekday and is_weekend

Label Encoding

For Category column that have 2 unique value or ordinal data

One Hot Encoding

For Category column that have more than 2 unique value or nominal data

Standar Scaler

For Numeric Column

Want to see the code? Click Here

Data Preprocessing

We split the data into 2 parts, there are features and target

We don't need to oversampling or undersampling the data because the target have balanced amount.

Want to see the code? Click Here

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Rakamin Academy

Modelling

For this modelling, we will do some experiment, there are :

1. First model will be trained by data that numeric columns dont do a scaling
2. Second model will be trained by data that has passed all data preprocessing

And in this case, I will try library called Lazypredict to make a model, and another will be i try are LightGBM, RandomForest and XGBoost.

Want to see the code? Click Here

Result of Experiment 1

Rakamin Academy

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
XGBClassifier	0.96	0.96	0.96	0.96	0.14
RandomForestClassifier	0.96	0.96	0.96	0.96	0.17
LGBMClassifier	0.95	0.95	0.95	0.95	0.17
BaggingClassifier	0.95	0.95	0.95	0.95	0.04
LogisticRegression	0.94	0.94	0.94	0.94	0.02
AdaBoostClassifier	0.94	0.94	0.94	0.94	0.12
LinearSVC	0.94	0.94	0.94	0.94	0.02
ExtraTreesClassifier	0.94	0.94	0.94	0.94	0.18
CalibratedClassifierCV	0.94	0.94	0.94	0.94	0.08
DecisionTreeClassifier	0.94	0.93	0.93	0.93	0.01
PassiveAggressiveClassifier	0.94	0.93	0.93	0.93	0.02
RidgeClassifierCV	0.94	0.93	0.93	0.93	0.04
RidgeClassifier	0.94	0.93	0.93	0.93	0.02
LinearDiscriminantAnalysis	0.94	0.93	0.93	0.93	0.03
NearestCentroid	0.94	0.93	0.93	0.93	0.01
Perceptron	0.93	0.93	0.93	0.93	0.02
SVC	0.93	0.92	0.92	0.92	0.03
SGDClassifier	0.92	0.92	0.92	0.92	0.02
NuSVC	0.92	0.92	0.92	0.92	0.05
BernoulliNB	0.91	0.90	0.90	0.90	0.01
ExtraTreeClassifier	0.83	0.83	0.83	0.83	0.02
GaussianNB	0.79	0.79	0.79	0.79	0.01
KNeighborsClassifier	0.71	0.71	0.71	0.71	0.03
LabelSpreading	0.70	0.70	0.70	0.70	0.08
LabelPropagation	0.70	0.70	0.70	0.70	0.08
DummyClassifier	0.48	0.50	0.50	0.32	0.01
QuadraticDiscriminantAnalysis	0.48	0.50	0.50	0.33	0.05

The top model is XGBClassifier from XGBoost model, there have high evalution score (like 96% accuracy), but the time to predict testing data is 0.14 (see time taken column), is 7x longer then logistic regression that have 94% accuracy.

So if you need model that have higher accuracy, you can use XGBClassifier but if you need model that have a faster time to predict you can use Logistic regression.

Want to see the code? Click Here

Result of Experiment 2

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
XGBClassifier	0.96	0.96	0.96	0.96	0.09
RandomForestClassifier	0.96	0.96	0.96	0.96	0.15
LGBMClassifier	0.95	0.95	0.95	0.95	0.17
BaggingClassifier	0.95	0.95	0.95	0.95	0.04
LogisticRegression	0.94	0.94	0.94	0.94	0.02
AdaBoostClassifier	0.94	0.94	0.94	0.94	0.23
LinearSVC	0.94	0.94	0.94	0.94	0.02
ExtraTreesClassifier	0.94	0.94	0.94	0.94	0.14
CalibratedClassifierCV	0.94	0.94	0.94	0.94	0.11
DecisionTreeClassifier	0.94	0.93	0.93	0.93	0.02
PassiveAggressiveClassifier	0.94	0.93	0.93	0.93	0.01
RidgeClassifierCV	0.94	0.93	0.93	0.93	0.03
RidgeClassifier	0.94	0.93	0.93	0.93	0.01
LinearDiscriminantAnalysis	0.94	0.93	0.93	0.93	0.03
NearestCentroid	0.94	0.93	0.93	0.93	0.01
Perceptron	0.93	0.93	0.93	0.93	0.01
SVC	0.93	0.92	0.92	0.92	0.03
SGDClassifier	0.92	0.92	0.92	0.92	0.01
NuSVC	0.92	0.92	0.92	0.92	0.04
BernoulliNB	0.91	0.90	0.90	0.90	0.01
ExtraTreeClassifier	0.83	0.83	0.83	0.83	0.02
GaussianNB	0.79	0.79	0.79	0.79	0.01
KNeighborsClassifier	0.71	0.71	0.71	0.71	0.02
LabelSpreading	0.70	0.70	0.70	0.70	0.05
LabelPropagation	0.70	0.70	0.70	0.70	0.05
QuadraticDiscriminantAnalysis	0.53	0.53	0.53	0.52	0.04
DummyClassifier	0.48	0.50	0.50	0.32	0.02

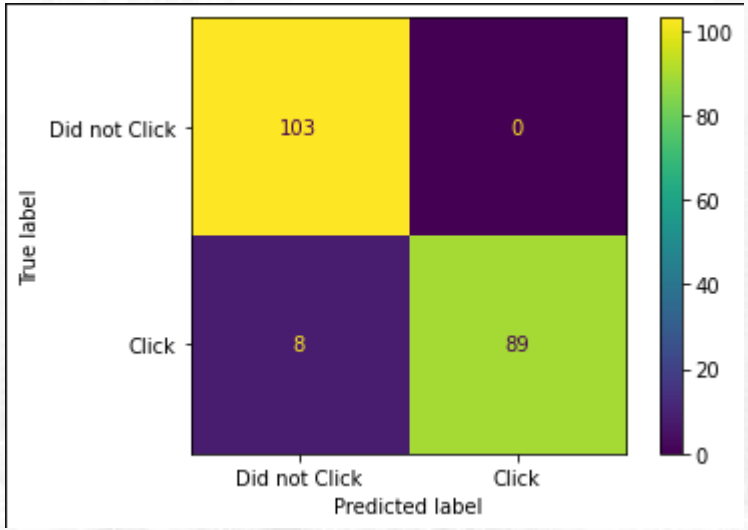
Compared with the result of experiment 1, there are no significant differences, just the time to predict test data is faster. We can look at XGBClassifier, at first experiment time taken is 0.14 second and at second experiment time taken is 0.09 second.

Want to see the code? Click Here

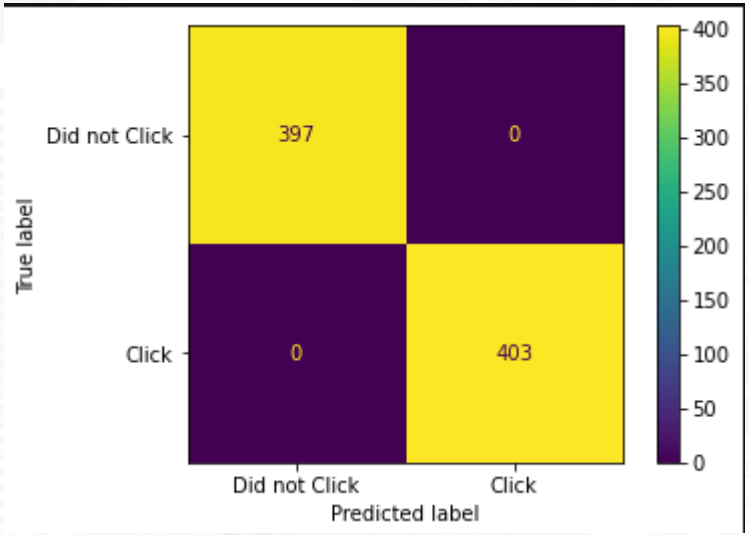
Model Evaluation

Rakamin Academy

Testing Data



Training Data



We can see our model at testing data made a detection error at data that customer was click our Ads. But the amount is few so I think we don't need to do a hyperparameter

Want to see the code? Click Here

Summary of 2 Experiment

1. The model is trained by preprocessing data to get better result.
2. XGBClassifier from XGBoost get the better evaluation score then another model. But if you need model that can predict data faster than XGBoost I recommend to use logistic regression.
3. Because I need model that have a better evaluation score, I will choose XGBClassifier.

Want to see the code? Click Here

Feature Importance

There are top 4 features that affect customers whether they click on our ads or not, namely Area income, Daily Internet Usage, Daily Time Spent on Site and Age.

Want to see the code? Click Here

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Business Recommendation

Business Recommendation

Based on Anaysis and feature importance from model, it can be concluded that :

1. We need to increase showing our ads to customer that meet the following requirement : They have income maximum 3.2 hundred million, are >=40 years old, have <= 60 minutes daily time spent on our site and have <= 170 minutes daily internet usage.

2. For customer that don’t meet criteria at number 1, we need to decrease showing our ads because they are have low amount customer to click our ads. So that we can maximize our budget in advertising.

Want to see the code? Click Here

Model Based - Simulation

We have a balanced amount of data between targets (50% click our ads and 50% no click our ads). Let's count if we don't use our model/do business recommendation :

Assumption :

We show our ads at Google searchs Ads that have average CPM $38.40 (at 2021 via topdraw.com), let's say if customer click our ads we got $0.1. so :

Cost : $38.40

Revenue = (1000 * 50%) * $0.1 = $50

Profit = Revenue – Cost = $50 - $38.40 = $11.6

Want to see the code? Click Here

Model Based - Simulation

Now if we use a ML model that has 96% accuracy to determine whether our customers will see our ads or not. We can get a profit per 1000 views of (We use the same assumption as before) :

Cost : $38.40
Revenue = (1000 * 96%) * $0.1 = $96
Profit = Revenue – Cost = $96 - $38.40 = $57.6

Is nearly 5 times bigger than if we don’t use ML model.

Want to see the code? Click Here