

PREDICT CUSTOMER CLICKED ADS CLASSIFICATION BY MACHINE LEARNING



Created by:
Putrini Nur Amalina H.
putrininur@gmail.com
[linkedin.com/in/putrininur](https://www.linkedin.com/in/putrininur)
github.com/putrini

A Diponegoro University graduate who experienced working in the Finance and Accounting Department at a Food Distribution company. A data-driven and tech-savvy person who has huge interest in data analytics who is skilled in SQL, Python, and data visualization using Google Data Studio. Highly skilled in Microsoft Excel and able to actively communicate in English fluently.

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

OVERVIEW

Machine learning is a method of data analysis that automates analytical model building. For organizations overflowing with data but struggling to turn it into useful insights, machine learning can provide the solution to analyze and make data-driven recommendations and decisions. Businesses use machine learning to recognize patterns and then make predictions about what will appeal to customers, improve operations, or help make a product better.

Classification is a machine learning algorithm of categorizing data into classes as labels or targets. By analyzing the input, the model learns how to classify new information and mapping labels or targets to the data. A use case of classification is often used in Banking, Financial Services, Insurance, etc. Finding fraud transactions and sophisticated classification algorithms is one of the business implementation on a real-time basis. In this paper we will predict the clicked ads customers using classification machine learning model.

GOAL

The goal is to predict the clicked ads customer based on customer profile and behavior to generate insights that lead to data driven business recommendation which objectives is to boost marketing campaign.

TOOLS



Python as
Programming language



Google colab as
notebook

EXPLORATORY DATA ANALYSIS

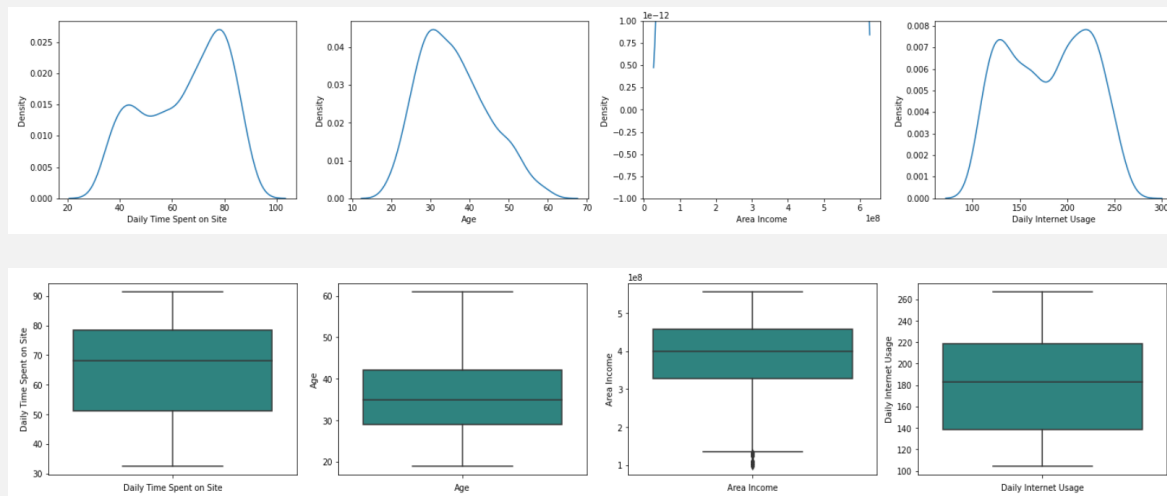
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 11 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   Unnamed: 0            1000 non-null   int64
1   Daily Time Spent on Site 987 non-null   float64
2   Age                   1000 non-null   int64
3   Area Income           987 non-null   float64
4   Daily Internet Usage   989 non-null   float64
5   Male                   997 non-null   object
6   Timestamp             1000 non-null   object
7   Clicked on Ad          1000 non-null   object
8   city                   1000 non-null   object
9   province               1000 non-null   object
10  category                1000 non-null   object
dtypes: float64(3), int64(2), object(6)
memory usage: 86.1+ KB
```

Insight :

- Data consist of 1000 rows and 11 columns.
- There are missing value in these columns :
 - Daily Time Spent on Site
 - Area Income
 - Daily Internet Usage
 - Male
- Data types in dataset are int64(2), float64(3) and object(6)

EXPLORATORY DATA ANALYSIS

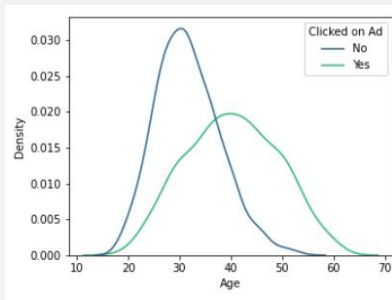
UNIVARIATE ANALYSIS



The distribution of all numerical columns are close to normal distribution and have no outliers detected

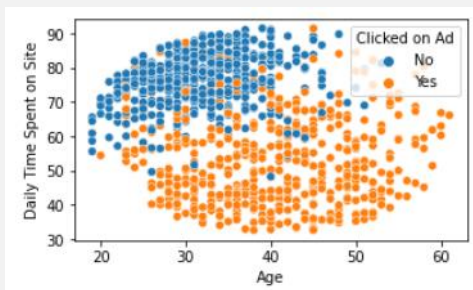
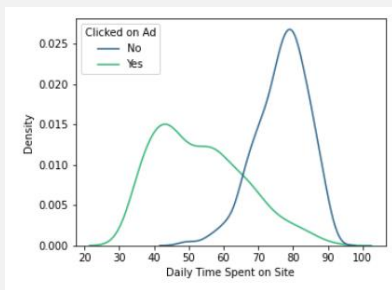
EXPLORATORY DATA ANALYSIS

BIVARIATE ANALYSIS



Age

Customers who clicked on ads have older age compared to user who don't clicked on ads. User with younger age are usually more susceptible to digital ads therefore they tend to avoid the feature.

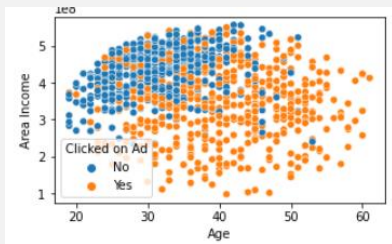


Daily Time Spent on Site

Customers who clicked on ads have fewer daily time spent on site rather than user who don't clicked on ad. Probably user with less time on the internet is more interested to ads.

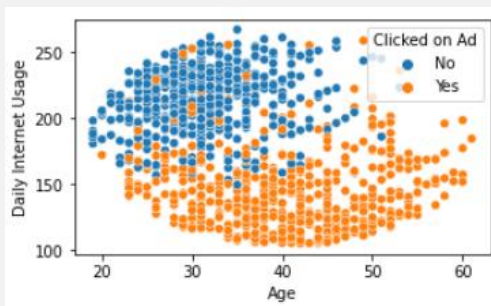
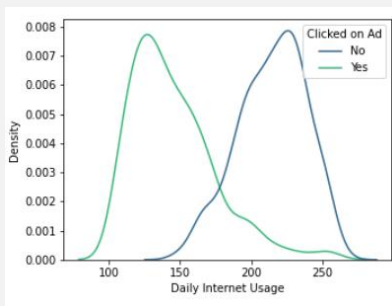
EXPLORATORY DATA ANALYSIS

BIVARIATE ANALYSIS



Area Income

Customers who clicked on ads are most likely to have fewer income and older age.

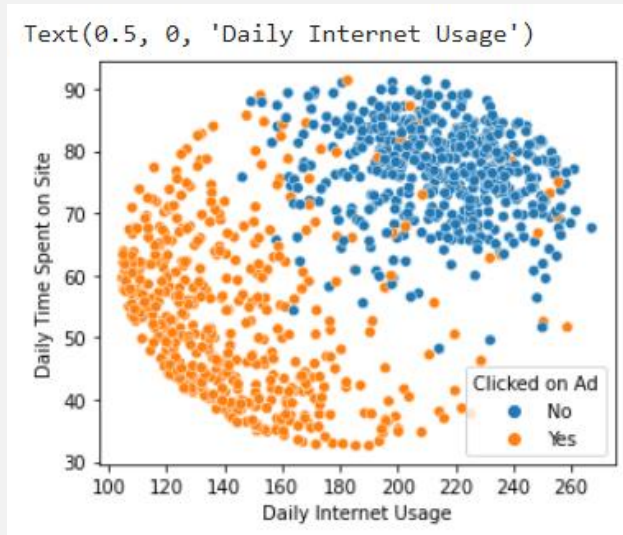


Daily Internet Usage

Customers who clicked on ads have fewer daily internet usage rather than user who don't click on ads. Same as the previous point, less internet usage tends to take less time on the internet which user with less internet time are probably more interested to digital ads.

EXPLORATORY DATA ANALYSIS

BIVARIATE ANALYSIS



Based on daily time spent on site and daily internet usage, users are segmented into 2 groups :

- Active user who have higher time spent on site and internet usage. This group have low potential to clicked on ad.
- Passive user who have lower time spent on site and internet usage. This group have high potential to clicked on ad.

Based on these insights, we can optimize our advertisement feature to passive user who have lower daily time spent on site and daily internet usage.

EXPLORATORY DATA ANALYSIS

MULTIVARIATE ANALYSIS



Based on PPS matrix above, the good predictors for Clicked on ad column as target are:

- Daily internet usage
- Daily time spent on site
- Age
- Area income

DATA PREPROCESSING

HANDLE NULL VALUES

Daily Time Spent on Site	13
Area Income	13
Daily Internet Usage	11
Male	3

How to handle :

- Daily time spent on site, Area income and Daily internet usage : imputation using mean based on age because data type is ratio and almost have normal distribution.
- Male = imputation using modus because data type is categorical.

HANDLE DUPLICATED VALUE

```
df.duplicated().sum()  
jumlah row duplicated 0
```

There is no duplicated values in dataset

FEATURE ENCODING

How to handle :

- City, Province, Category columns using one hot encoding

```
# handle dengan one hot encoding  
for cat in ['city', 'province', 'category']:  
    onehots = pd.get_dummies(df[cat], prefix=cat)  
    df = df.join(onehots)
```

- Male, Clicked on ad columns using label encoding

```
# label encoder  
mapping_male = {  
    'Laki-Laki' : 0,  
    'Perempuan' : 1}  
  
mapping_clicked = {  
    'No' : 0,  
    'Yes' : 1}
```

DATA PREPROCESSING

FEATURE ENGINEERING

Converting Timestamp column into year, month, week and day columns

```
df['year'] = pd.DatetimeIndex(df['Timestamp']).year  
df['month'] = pd.DatetimeIndex(df['Timestamp']).month  
df['day'] = pd.DatetimeIndex(df['Timestamp']).day  
df['week'] = pd.DatetimeIndex(df['Timestamp']).week
```

FEATURE SELECTION

Drop columns unused for further data modeling which are :

- Unnamed:0 = unique value
- Timestamp, Clicked on ad, Male, Province, City, Category = columns already converted into another feature

SPLIT TARGET AND FEATURES

Split columns into target and features dataset

```
X = df2.drop(labels=['adclicked_mapped'],axis=1)  
y = df2[['adclicked_mapped']]
```

DATA MODELING

1. Split train and test set

```
from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test = train_test_split(X,y,test_size = 0.3,stratify=y,random_state = 123)
```

2. Modeling (without Normalization)

	model_name	model	accuracy	recall	precision	duration
0	K-Nearest Neighbor	KNeighborsClassifier()	0.683333	0.606667	0.716535	0.008136
1	Logistic Regression	LogisticRegression()	0.500000	0.000000	0.000000	0.010815
2	Decision Tree	DecisionTreeClassifier()	0.943333	0.926667	0.958621	0.010684
3	Random Forest	(DecisionTreeClassifier(max_features='auto', r...	0.936667	0.940000	0.933775	0.274779
4	Gradient Boosting	([DecisionTreeRegressor(criterion='friedman_ms...	0.940000	0.933333	0.945946	0.239675

Based on modeling process, model with the highest accuracy score is Random forest, Decision tree, Gradient boosting and Random forest. But considering the duration, decision tree is the model with lowest duration time. These are modeling results without handling the data normalization and these score can be increased more through implementing normalization to the dataset.

DATA MODELING

3. Modeling with data normalization

- Data normalization using MinMaxScaler

	model_name	model	accuracy	recall	precision	duration
0	K-Nearest Neighbor	KNeighborsClassifier()	0.753333	0.700000	0.783582	0.007395
1	Logistic Regression	LogisticRegression()	0.940000	0.900000	0.978261	0.019050
2	Decision Tree	DecisionTreeClassifier()	0.940000	0.926667	0.952055	0.007378
3	Random Forest	(DecisionTreeClassifier(max_features='auto', r...	0.943333	0.926667	0.958621	0.266748
4	Gradient Boosting	([DecisionTreeRegressor(criterion='friedman_ms...	0.936667	0.933333	0.939597	0.266462

After handling data normalization, there are several changes to the model result. The accuracy has increased for all models especially in KNN and Logistic Regression who have significant increase compared to the previous result. This proved that handle data normalization can improve model performance including accuracy, recall and precision score.

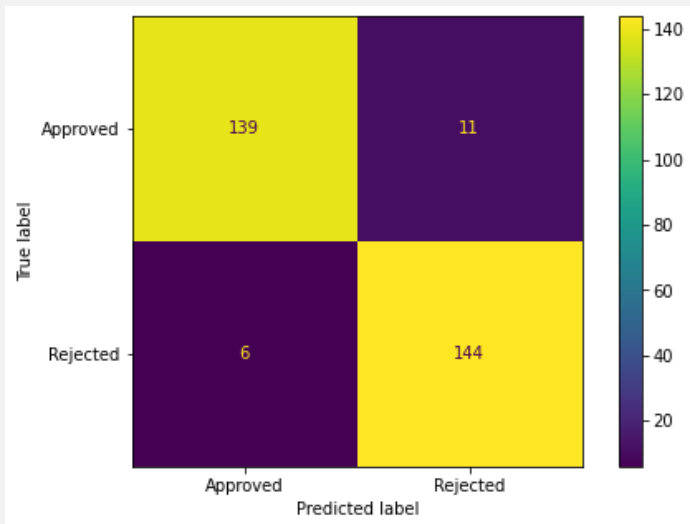
The objectives of our model is to predict the maximum amount of potential customer who clicked on ads. Therefore, we should minimize the False Positive where customers who don't clicked on ad are falsely predicted to be clicked on ad. This will lead to wrong market targeting then lead to potential loss because of the false target on marketing cost spent.

Therefore, **we have to optimize the precision score while still considering other metrics to be in maximum score.**

Finally, **random forest model is chosen considering the high precision and accuracy score.**

EVALUATION

Confusion Matrix

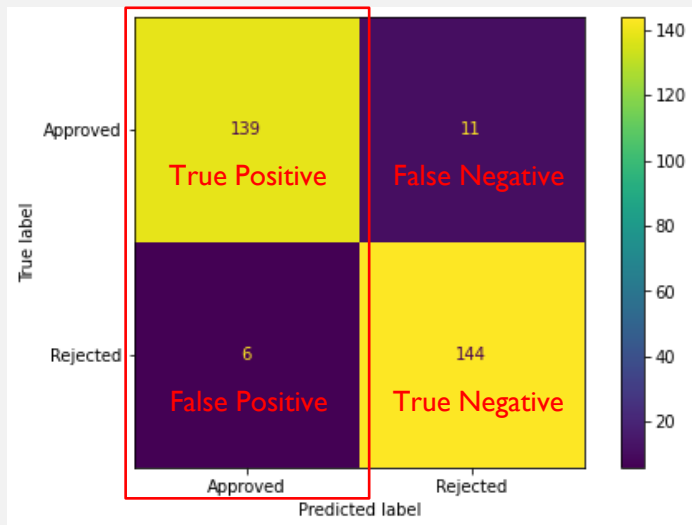


Dataset divided into classes :

- **True Positive** is the number of observation where model predicted the customer would click on ad and they actually clicked on ad (139 customers)
- **False Negative** is the number of observation where model predicted the customer would not click on ad, but they actually clicked on ad (11 customers)
- **False Positive** is the number of observation where model predicted the customer would click on ad, but they actually not clicked on ad (6 customers)
- **True Negative** is the number of observation where model predicted the customer would not click on ad, and they actually not clicked on ad (144 customers)

EVALUATION

Confusion Matrix



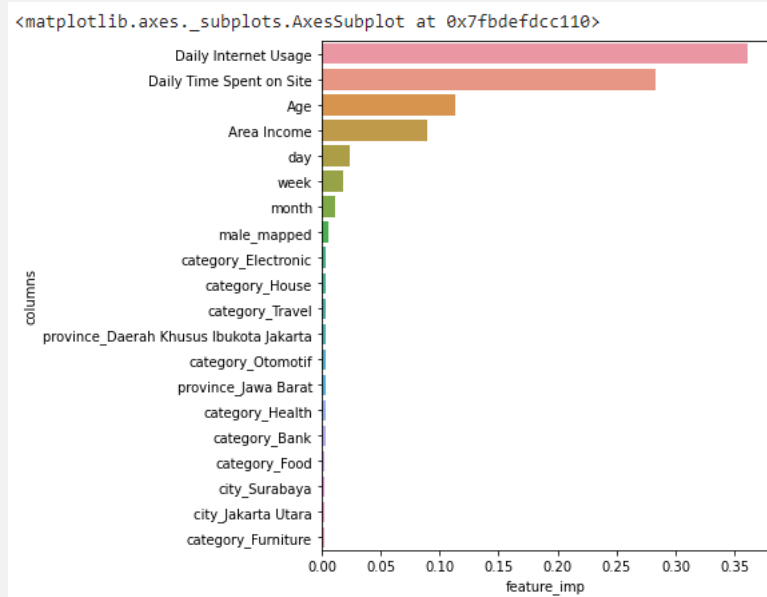
- The objectives of our model is to predict the maximum amount of potential customer who clicked on ads.
- Therefore, we should minimize the False Positive where customers who don't clicked on ad are falsely predicted to be clicked on ad. Because this will lead to wrong market targeting then lead to potential loss because we have spent marketing cost on false target.

Based on the confusion matrix, our model have generated great result where the number of False Positive (Predicted clicked ad, but actually not) has optimally minimized at 2% rate.

This leads to high True Positive score (Predicted clicked ad and actually did) at 46,3% where leads to higher potential profit.

EVALUATION

Feature Importance



Based on feature importance graphs, we can see that there are 4 features which have high correlation in building the model, there are =

- Daily Internet Usage
- Daily Time Spent on Site
- Age
- Area Income

Feature with the highest correlation in predicting clicked ads customers is Daily Internet Usage and Daily Time Spent on Site. This is proven based on previous EDA analysis that customers with lower daily internet usage and time spent on site have higher potential to clicked on ads.

Other features which also have high feature importance are age and area income. This supported by customers distribution mapping based on age and area income shows that customers who have younger age and higher income have low potential to clicked on ads.

BUSINESS RECOMMENDATION

BASED ON EDA AND FEATURE IMPORTANCE ANALYSIS WE CAN SUMMARIZE :

Customer Classification

We can divide customers based on customers characteristics and behavior.



High class customers

These customers characterized by lower age and higher income. They usually spend higher daily internet usage and more time on internet

- High class customers who have higher daily internet usage and more time on internet are less likely to clicked on ads.
- This may happen because these customers are already used to digital ads, so they tend to avoid ads and clicked them.



Low class customers


These customers characterized by older age and lower income. They usually spend lower daily internet usage and fewer time on internet.

- Low class customers who have few daily internet usage and few time on internet are tend to more clicked on ads.
- This is surprisingly shows that customers with low exposure to internet are more interested into digital ads.
- Customers with older age are high potential market target.
- We can try to include the digital ads content to website where usually surfed by people with older ages such as social media platform for ex. Facebook, etc.

BUSINESS RECOMMENDATION

Rocket Songs
February 9 at 11:32am · 🌐

License original songs by the industry's best songwriters at RocketSongs. We've got licensing options to match every goal and budget
<http://ow.ly/XBixX>



Original Songs for Original Artists

Pick a Genre, Any Genre

15,067 people reached

Like Comment Share

Mi Kattar, Kyaw Kyaw, Mg Mg and 737 others like this.

Starbucks
February 6 · 🌐

Join us for the World's Largest #StarbucksDate! February 13th from 2 P.M. to close.



#StarbucksDate
Explore our perfect pairings for this lovely occasion: French press Caffé Verona & a chocolate brownie, Raspberry or White Chocolate Mocha & a Heart Cookie, or...

ATSTARBUCKS.TUMBLR.COM

Like · Comment · Share · 40,345 1,530 3,657

We can **target the digital ads to website where most likely to be visited by older age customers.**

Potential websites are social media such as Facebook and entertainment/video streaming website such as YouTube, etc.

BUSINESS RECOMMENDATION



Implement the soft selling marketing that shows less likely like an ad and more integrated with the site content.

This is to targeting the high-class customers who have more time spent on internet.

BUSINESS RECOMMENDATION



Develop digital ads using mainstream content with simple but cooperated with topic that are trending recently.

SIMULATION

Assumptions

- Marketing cost per customer = Rp 10.000
- Profit gained per customer who clicked on ads = Rp 15.000
- We will do the simulation targeting to population of 300 customers

WITHOUT MACHINE LEARNING

Based on customer distribution in clicked on ads label, customers are divided into two groups with 50 : 50 probability

There are customers who clicked on ads (150 customers) and customers who don't click on ads (150 customers)

```
No      500  
Yes      500  
Name: Clicked on Ad, dtype: int64
```

Marketing cost

300 customers x Rp 10.000 = Rp 3.000.000

Conversion rate

Clicked on ads customer / total customers = $150/300 \times 100\% = 50\%$

Revenue

150 customers x Rp 15.000 = Rp 2.250.000

Profit

Revenue - Cost = - Rp 750.000

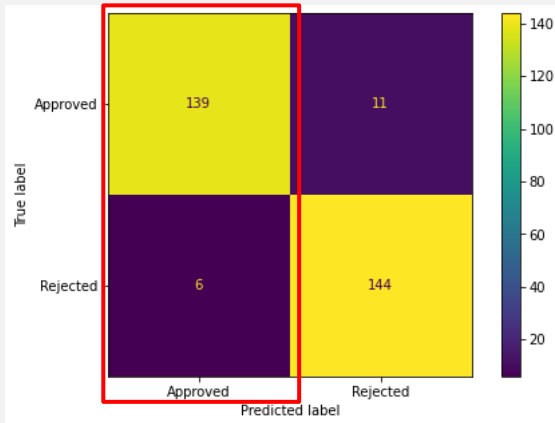
Based on simulation above, **the amount of potential loss is Rp 750.000** with **loss rate around 25%.**

SIMULATION

Assumptions

- Marketing cost per customer = Rp 10.000
- Profit gained per customer who clicked on ads = Rp 15.000
- We will do the simulation targeting to population of 300 customers

WITH MACHINE LEARNING



Based on confusion matrix, we can classify the customers based on the prediction who will clicked on ads (144 customers) and who don't click on ads (156 customers).

Marketing cost

144 customers X Rp 10.000 = Rp 1.450.000 ▼

Conversion rate

Clicked on ads customer / total customers = $139/145 \times 100\% = 95.86\%$ ▲

Revenue

139 customers X Rp 15.000 = Rp 2.085.000 ▲

Profit

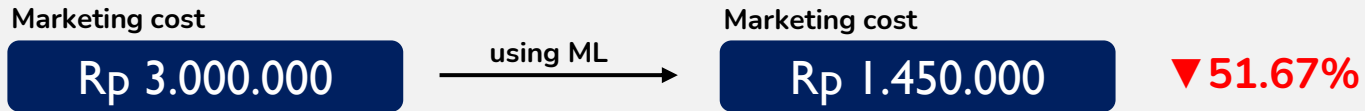
Revenue - Cost = Rp 635.000 ▲

Based on simulation above, the amount of profit is Rp 635.000 with profit rate 43,8%.

CONCLUSION

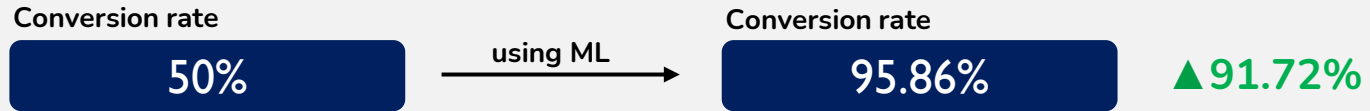
Based on simulation, we can summarize that :

- The amount of marketing cost invested in simulation without machine learning is higher and have no clear target since we spent the cost to the entire populations. This will lead to inefficient used of marketing cost and higher potential loss without any profit from its revenue.
- Meanwhile the amount of marketing cost invested in simulation implementing ML model is fewer because we already have prediction on customers who clicked on ads and who don't click on ads among the population of 300 customers. Therefore, we will focus on group with predicted to clicked on ads customers.

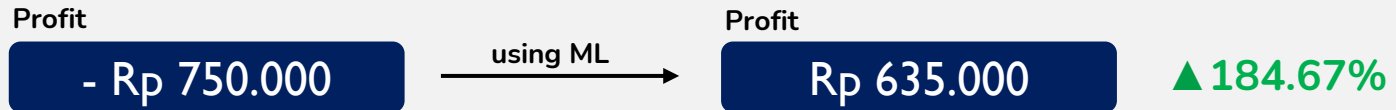


CONCLUSION

- By implementing machine learning model, not only we will decrease the number of marketing target but also increasing the efficiency because we invested the marketing cost on the right population. This leads to higher conversion rate up to 96%.



- Then we can see the model able to increase the profit rate up to 186% compared to the previous simulation without machine learning model implementation.



- Thus, the implementation of machine learning in predicting clicked ads customers is useful in the industry because this will prevent business to experiences potential loss and leads to higher potential profit.