Predict Clicked Ads Customer Classification by using Machine Learning

Created by:
Dinda Galuh
dindagaluhg@gmail.com
linkedin.com/in/dinda-galuh-guminta

Statistics graduate with passion in data visualization, data analysis, and reporting. Dedicated and hard working person.

I joined the data science bootcamp because I am more interested and have a passion for learning and working in the data field. I like doing data analysis and data visualization. I want to apply my knowledge in the real world.

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

Business Understanding

Problem Statement:

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 find out how much the advertising has been marketed so that it can attract customers to see the advertisement.

By processing historical advertisement 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.

Goals:

Increase the **effectiveness of an advertisement** by targeting the right customers so that **response rates and profits increase**.

Objective:

Predict Clicked Ads Customer Classification

Business Metrics:

- Click rate
- Profit

Dataset Overview

Daily Time Spent on Site	: consumer time on site in minutes
Age	: customer age in years
Area Income	: Avg. Income of geographical area of consumer
Daily Internet Usage	: Avg. minutes a day consumer is on the internet
Ad Topic Line	: Headline of the advertisement
City	: City of consumer
Male	: Whether or not consumer was male
Country	: Country of consumer
Timestamp	: Time at which consumer clicked on Ad or closed window
Clicked on Ad	: 0 or 1 indicated clicking on Ad

EDA

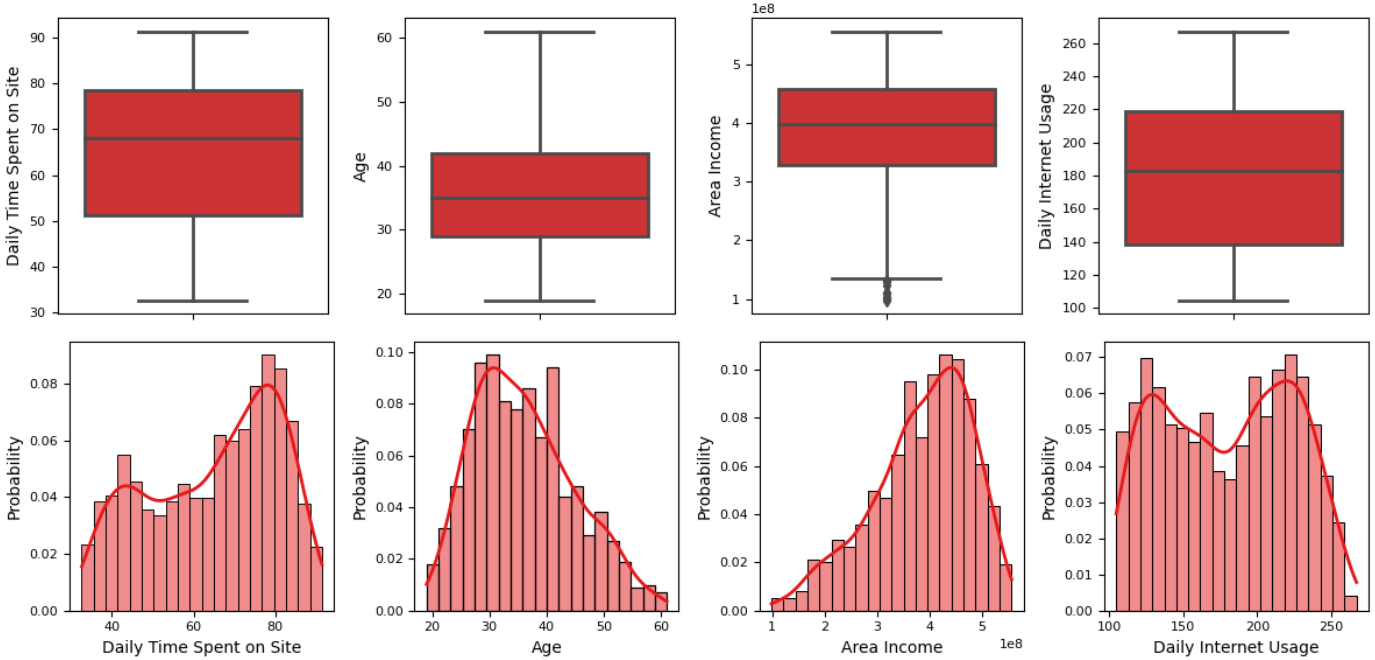
- Descriptive Statistics
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis

Descriptive Statistics

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage
count	987.000000	1000.000000	9.870000e+02	989.000000
mean	64.929524	36.009000	3.848647e+08	179.863620
std	15.844699	8.785562	9.407999e+07	43.870142
min	32.600000	19.000000	9.797550e+07	104.780000
25%	51.270000	29.000000	3.286330e+08	138.710000
50%	68.110000	35.000000	3.990683e+08	182.650000
75%	78.460000	42.000000	4.583554e+08	218.790000
max	91.430000	61.000000	5.563936e+08	267.010000

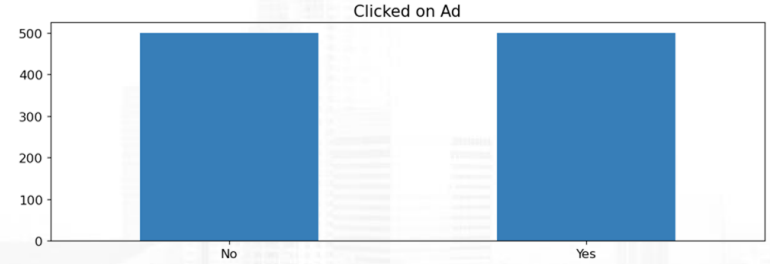
- There are no duplicate values in the data
- The variables age and income area have almost the same mean and median, so the data almost has a normal distribution pattern
- The daily time spent on site and daily internet usage variables have a mean that is smaller than the median which indicates a negatively skewed distribution pattern
- There are missing values in the variables daily time spent on site, income area, and daily internet usage

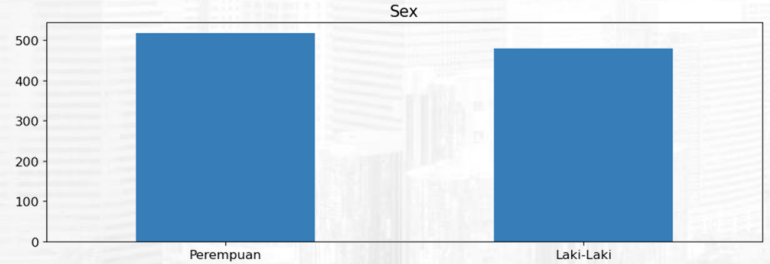
Univariate Analysis

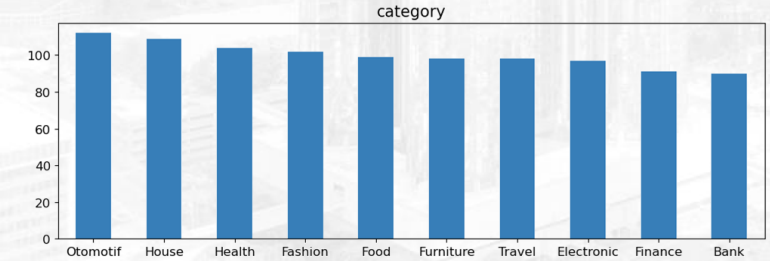


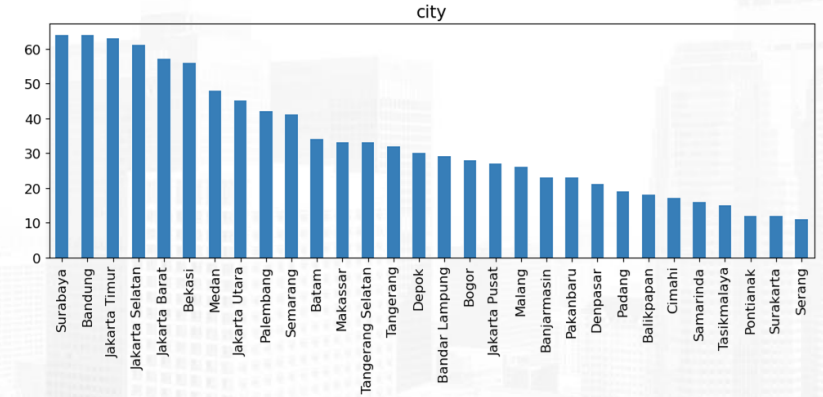
- Age has a positively skew distribution pattern (skewed to the right) where the mean > median
- The income area has a negatively skewed distribution pattern (skewed to the left) where the mean < median, you can see the outlier value on the boxplot which causes the tail to elongate on the left side of the histogram
- The daily time spent on site has a negatively skew pattern but there is a bimodal appearance, namely two peaks in the variable which indicates that the data indicates that there are two different groups in the population
- Daily internet usage shows a bimodal distribution pattern.

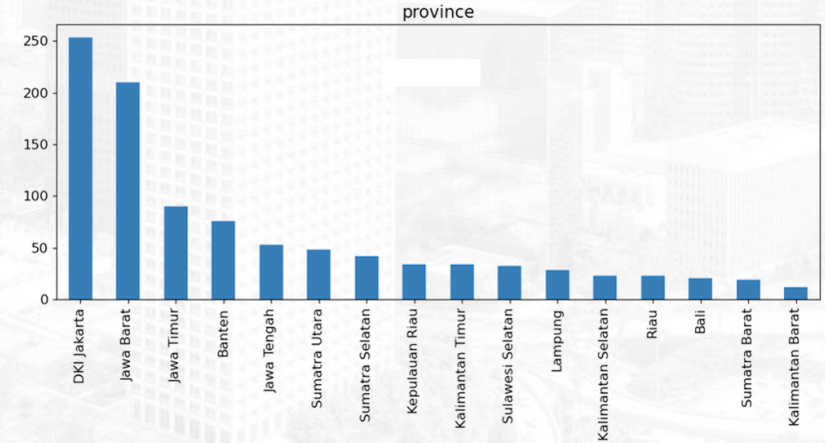
Univariate Analysis





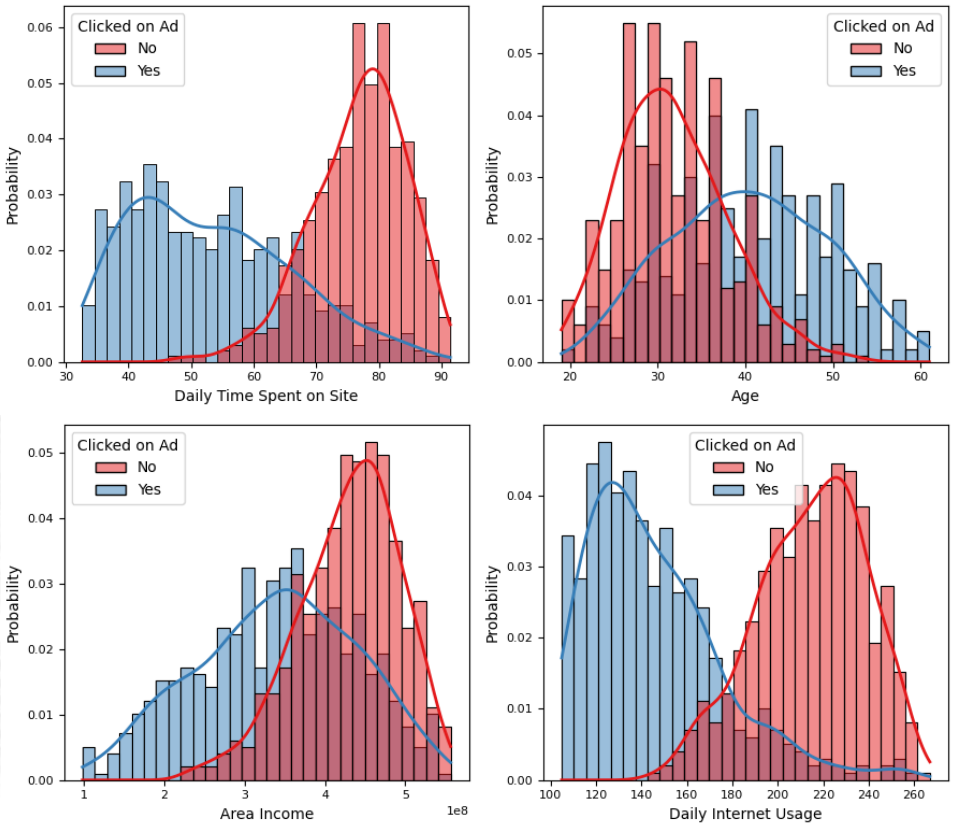






- Gender, click on ad, and category variables have almost equal proportions in each category
- The categories for city and province variables have unequal proportions

Bivariate Analysis with Click on Ad



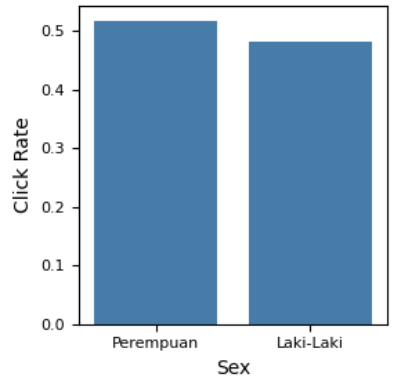
Insight:

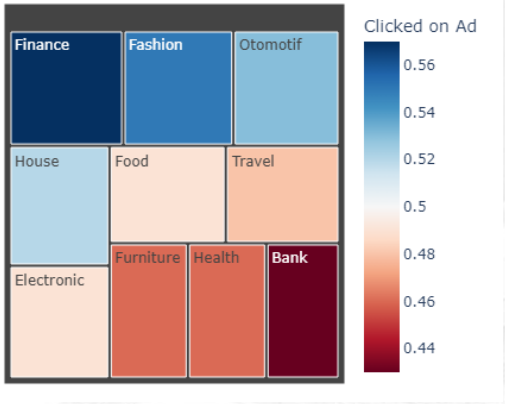
- users who did not click the ads spent an average of longer time on the site (about 76 minutes on average) than those who did click the ads (about 45 minutes on average). This suggests that the length of time a user spends on a site has a relationship with the likelihood that they will click on an ad.
- the average age of users who did not click on ads was around 30 years, while users who did click on ads had a higher age of 40 years.
- users who do not click ads have higher average daily internet usage and income than users who click ads.

Business Recommendation:

1. **Short Session Duration:** this can be done by placing the ad in a more prominent place to encourage them to click the ad earlier in the session.
2. **Optimize Ads:** placing ads in areas of the site that users visit most in their initial session. Ad Optimization for Users with High Internet Usage.
3. **Target Demographics & Ads Personalization:** display ads that are relevant to the needs and interests of users in each particular age and income group as well as in groups with low Internet usage.

Bivariate Analysis – Click Rate by Sex and Category





Insight:

- Higher click rates among women than men.
- Click rates vary by product category. 'Finance' category products had the highest click rate (57%), followed by 'Fashion' (55%), 'Automotive' (53%) and 'House' (52%). On the other hand, the 'Bank' category had the lowest click rate (43%), followed by 'Furniture' and 'Health' (both 46%).

Business Recommendation:

1. **Ad Targeting:**increased ad targeting for products in the 'Finance', 'Fashion', 'Automotive' and 'House' categories.Ads can be adapted to appeal more to women or products that are more popular with women can be promoted more intensively.
2. **Ad Optimization:** For categories with lower click rates like 'Bank', 'Furniture' and 'Health', you may need to have more attractive ads, offer promotions or discounts, or perhaps try a different marketing approach.
3. **Ad Personalization:** Because women and men have different click rates, this personalization can involve adjusting ad content, bidding, and ad serving time.Increasing Male Engagement: Specific marketing and advertising strategies to attract males.

Bivariate Analysis – Click Rate by City and Province

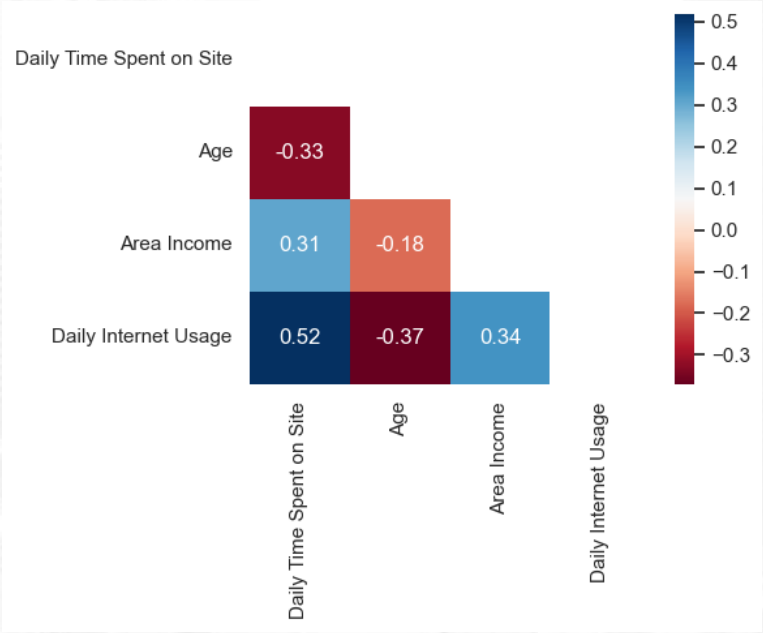
Insight:

Ad click rates vary depending on city and province. 'Serang' in 'Banten' province had the highest click rate (82%), followed by 'Cimahi' in 'West Java' (76%), and 'Tangerang Selatan' in 'Banten' (64%). On the other hand, 'Central Jakarta' in 'DKI Jakarta' had the lowest click rate (26%), followed by 'Balikpapan' in 'East Kalimantan' (28%) and 'Malang' in 'East Java' (31%).

Business Recommendation:

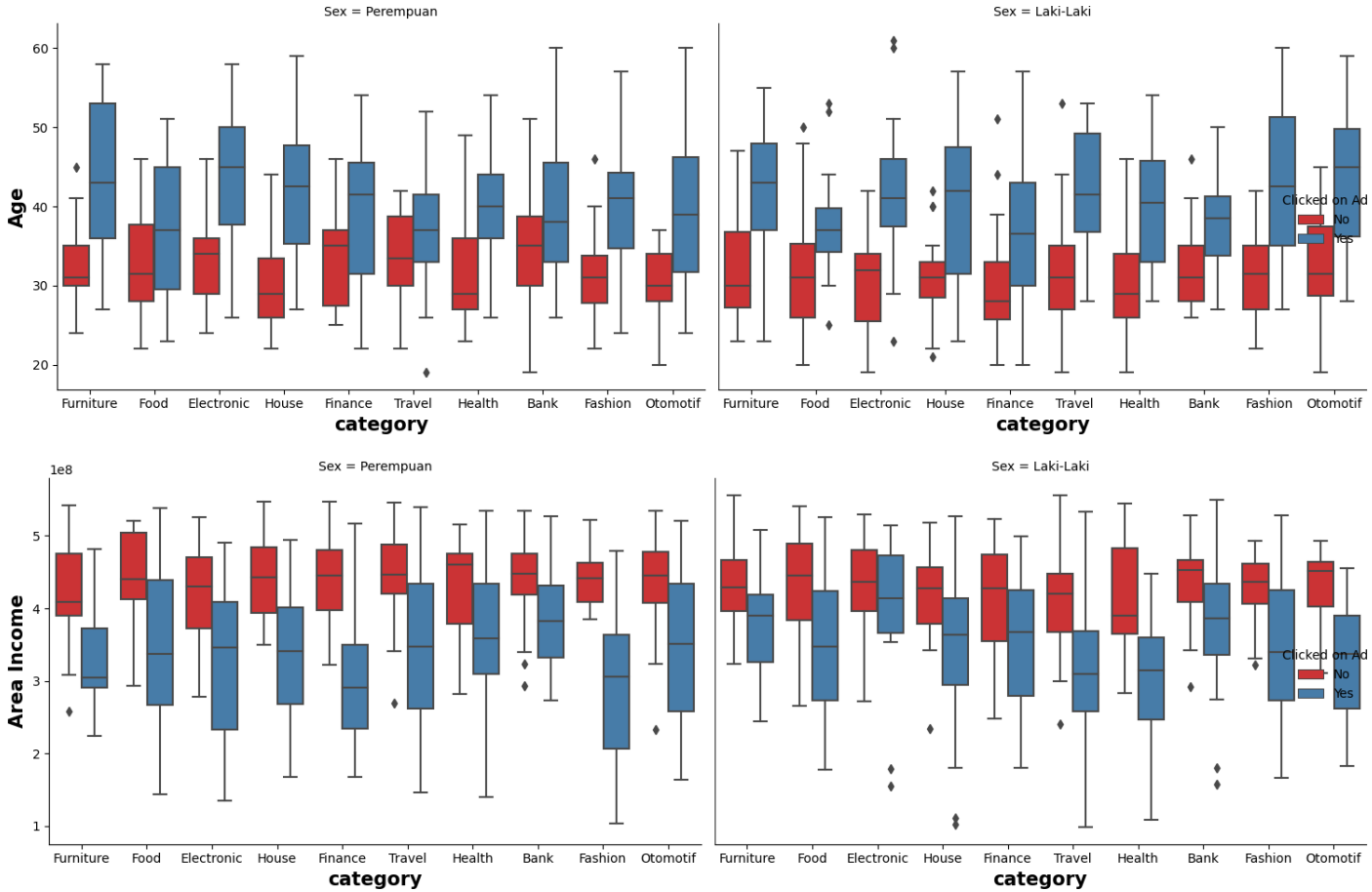
- **Geotargeting Ads in areas** such as 'Serang', 'Cimahi', and 'Tangerang Selatan'.
- **Increase Engagement in Areas with Low Click Rates** such as 'Central Jakarta', 'Balikpapan', and 'Malang', by creating more attractive ads, special offers, or adjusting ad serving time.

Multivariate Analysis : Correlation Analysis in Numerical Variables



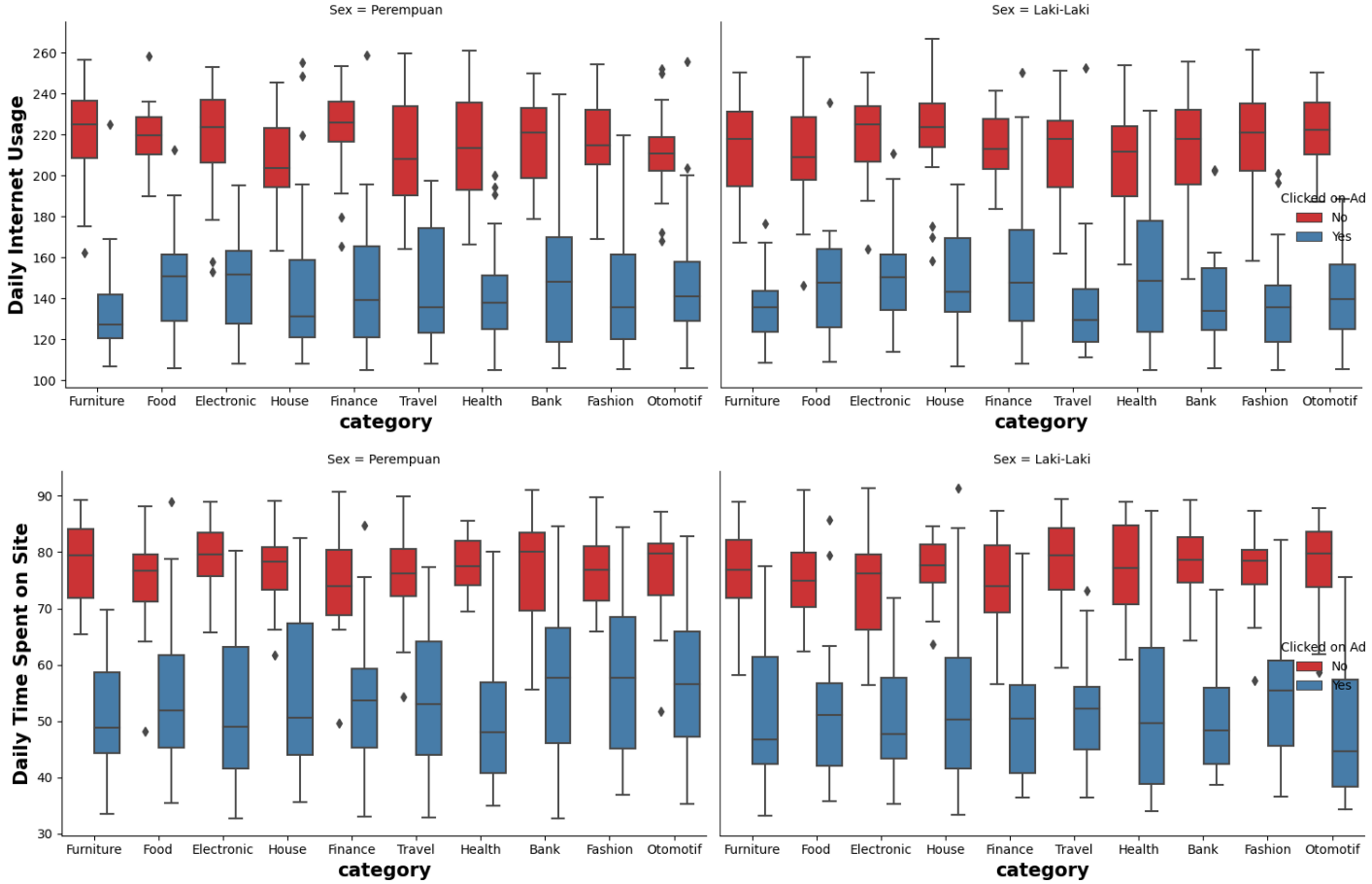
- Daily internet use and daily time spent on the site have a high correlation of 0.52 which indicates that the higher the daily time spent on the site, the greater the daily internet use.
- The more income, the more daily time spent on the site and the greater the daily internet usage with a correlation of 0.31 and 0.34.
- However, as age increases, the daily time spent on the site and daily internet use decreases because it has a negative correlation.
- The older you get, the smaller the value of income.

Multivariate Analysis : Age and Area Income by Sex, Category, and Clicked on Ad



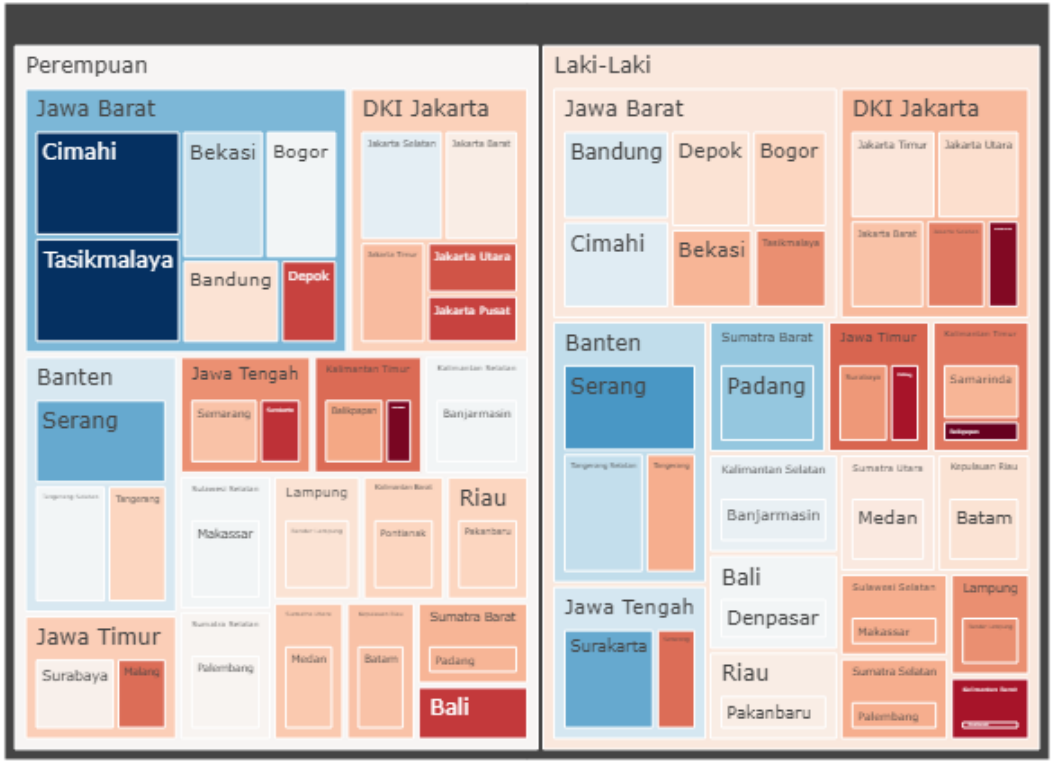
It appears that customers who click on ads have a higher average age and lower income than those who do not click on ads for both men and women.

Multivariate Analysis : Daily Internet Usage and Daily Time Spent on Site by Sex, Category, and Clicked on Ad



The distribution of data on daily internet usage and time spent on the internet for customers who click on ads is lower than for customers who don't click on ads for both men and women.

Multivariate Analysis : Click Rate by Sex, Province, and City



Ad click rates vary depending on gender, city and province.

- For female customers, the highest click rates were in the cities of Cimahi and Tasikmalaya in West Java, while the lowest click rates were in Balikpapan, East Kalimantan.
- For male customers, the highest click rates are in Serang City, Banten and Surakarta City, Central Java, while the lowest click rates are in Central Jakarta, Balikpapan City - East Kalimantan and Malang - East Java.

DATA PREPROCESSING

- Data Cleaning
- Split Data
- Feature encoding

Data Cleaning & Data Preprocessing

Rakamin Academy

Data Cleaning

Change Data Type

Split data

Feature Encoding

```
Daily Time Spent on Site    0.013
Area Income                 0.013
Daily Internet Usage        0.011
Sex                         0.003
```

```
df.duplicated().sum()

0
```

- Missing value pada kolom daily time spent on site dan area income diisi dengan nilai median, missing value pada kolom daily internet usage diisi dengan median pada masing-masing grup, menghapus baris dengan missing value pada kolom Sex
- Tidak terdapat baris yang duplikat
- Merubah tipe data pada kolom Timestamp
- Merubah nama kolom Male menjadi Sex

```
df['Timestamp'] = pd.to_datetime(df['Timestamp'])
```

```
df.rename(columns={"Male": "Sex"}, inplace=True)
```

```
df_prep['Timestamp'] = pd.to_datetime(df_prep['Timestamp'])
```

```
# Buat kolom baru untuk tahun, bulan, minggu dan hari
df_prep['Year'] = df_prep['Timestamp'].dt.year
df_prep['Month'] = df_prep['Timestamp'].dt.month
df_prep['Week'] = df_prep['Timestamp'].dt.week
df_prep['Day'] =df_prep['Timestamp'].dt.day
```

```
print(X_train.shape)    (697, 9)
print(y_train.shape)    (697, 1)
print('')
print(X_test.shape)     (300, 9)
print(y_test.shape)     (300, 1)
```

- Split data dengan ratio 70:30 dengan Click on Ads sebagai targer

- Melakukan one hot encoding pada variable Sex, city, province, category

Total row

Before preprocessing : 1,000

After preprocessing : 997

Modeling

In this analysis, 2 experiments were carried out

- Data without scaling
- Data with scaling

Model Evaluation Without Scaling Data

Rakamin Academy

Classifier	Accuracy		Precision		Recall		F1 Score		ROC AUC	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Decision Tree	1.0000	0.9467	1.0000	0.9408	1.0000	0.9533	1.0000	0.9470	1.0000	0.9467
Gradient Boosting	1.0000	0.9567	1.0000	0.9597	1.0000	0.9533	1.0000	0.9565	1.0000	0.9888
K-Nearest Neighbours	0.7661	0.6767	0.8000	0.6853	0.7106	0.6533	0.7527	0.6689	0.8462	0.7348
Logistic Regression	0.4993	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7576	0.7966
Random Forest	1.0000	0.9533	1.0000	0.9474	1.0000	0.9600	1.0000	0.9536	1.0000	0.9864

Decision Tree:

- **Training**: Model has achieved perfect scores on the training data for all metrics indicating 100% correct predictions. This is a strong sign of overfitting as decision trees tend to memorize the training data.
- **Testing**: The Decision Tree performs well on the test data, but not as perfect as the training data.

Gradient Boosting:

- **Training**: Like the Decision Tree, Gradient Boosting also shows perfect scores on the training data.
- **Testing**: The model achieves very high scores on the test data, but again not as high as the training data. However, the performance drop is not very significant, indicating that this model generalizes well.

K-Nearest Neighbors (KNN):

- **Training**: KNN achieves decent but not perfect scores, suggesting that the model might be underfitting or that the feature space isn't suitable for a distance-based method like KNN.
- **Testing**: The drop in performance in the test data.

Logistic Regression:

- **Training – Testing**: Logistic Regression achieves an accuracy close to 50% and zero precision, recall, and F1 score. This suggests that the model might be predicting only one class, possibly due to the **features have different scales or magnitudes**.

Random Forest:

- **Training**: Random Forest shows perfect scores on the training data, suggesting potential overfitting.
- **Testing**: The scores drop on the test data but still remain high. Random Forest typically handles overfitting better than a singular decision tree due to its ensemble nature.

Model Evaluation Using Scaling Data

Rakamin Academy

Classifier	Accuracy		Precision		Recall		F1 Score		ROC AUC	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Decision Tree	1.0000	0.9467	1.0000	0.9408	1.0000	0.9533	1.0000	0.9470	1.0000	0.9467
Gradient Boosting	1.0000	0.9567	1.0000	0.9597	1.0000	0.9533	1.0000	0.9565	1.0000	0.9888
K-Nearest Neighbours	0.9598	0.9567	0.9878	0.9858	0.9312	0.9267	0.9587	0.9553	0.9957	0.9852
Logistic Regression	0.9684	0.9733	0.9685	0.9863	0.9685	0.9600	0.9685	0.9730	0.9953	0.9871
Random Forest	1.0000	0.9533	1.0000	0.9474	1.0000	0.9600	1.0000	0.9536	1.0000	0.9864

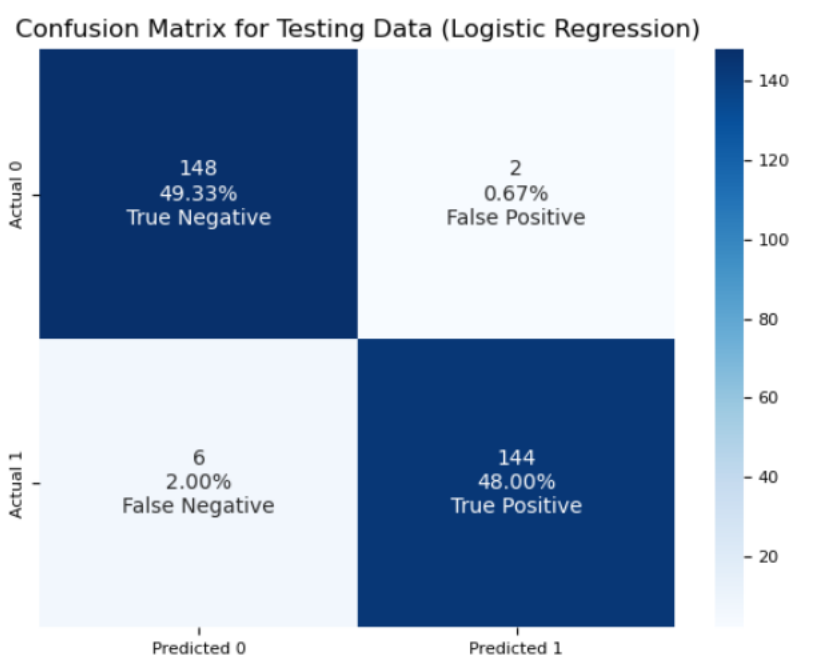
It can be seen that there is an **increase in performance in the logistic regression model and KNN**. This is because:

- In logistic regression, In unscaled data, features with larger numeric ranges will have smaller coefficients, while features with smaller numeric ranges will have larger coefficients. This is because a small change in a large-scale feature can have a much larger effect on the output than a small change in a small-scale feature.
- KNN relies on distance to find the closest neighbors. Scaling ensures that all features contribute equally to the distance computation, allowing the algorithm to make more informed and accurate predictions.

However, in models such as **decision trees, gradient boosting, random forests, the results are the same for data with or without scaling**. This is because:

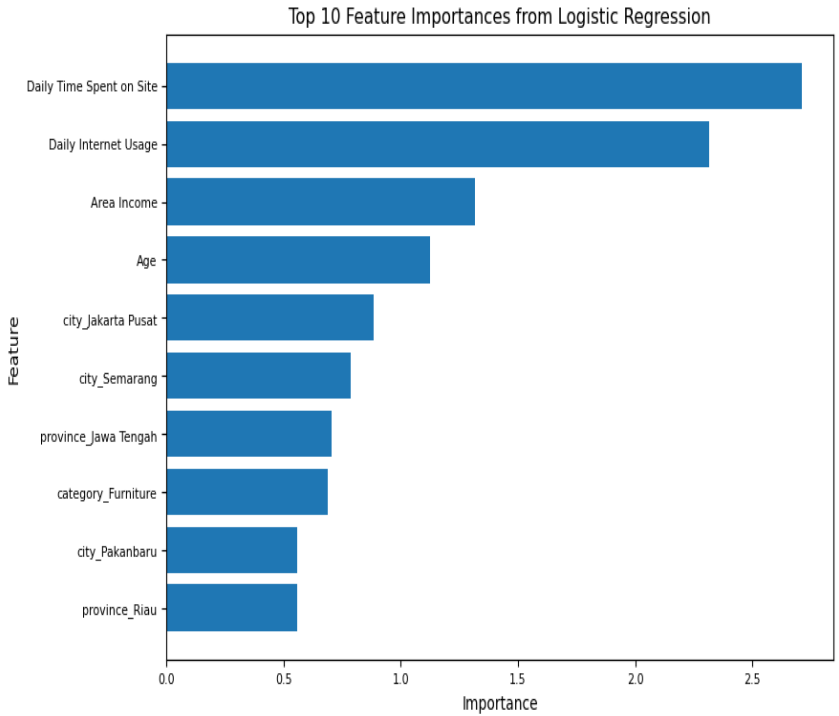
- Tree-based methods can capture non-linear relationships without requiring transformations or scaling, as they split on feature values rather than relying on distances or angles between data points

Confusion Matrix – Logistic Regression with Scaling Data



- The model has done a relatively good job in predicting users who would and would not click on the ad, with a small percentage of errors in both the false positive and false negative categories.
- The business impact of these errors would depend on the specific context. For example, **false positives** might be seen as wasted opportunities in some ad campaigns where a user is predicted to click but doesn't, especially in pay-per-click models. On the other hand, **false negatives** could represent missed opportunities for retargeting or further engagement since these are users who showed genuine interest.

Feature Importance – Logistic Regression with Scaling Data

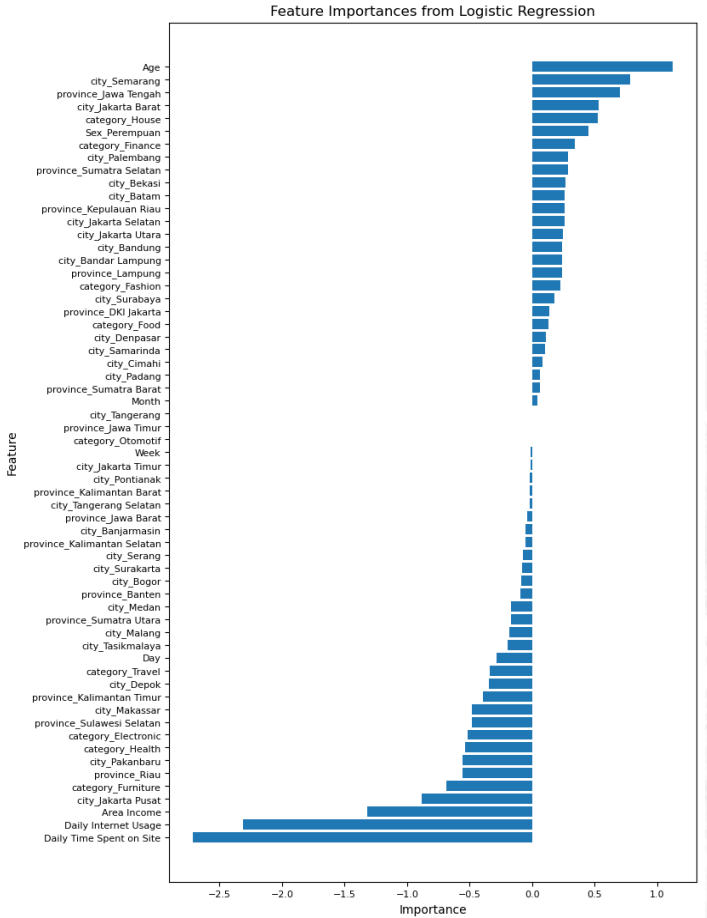


- "Daily Time Spent on Site" can be a direct indicator of user engagement. Users who spend more time on a site are more likely to be engaged and possibly more receptive to the content, including ads. They have a higher chance of noticing and clicking on ads compared to someone who just briefly visits the site.
- High "Daily Internet Usage" might indicate that a user is tech-savvy, familiar with the digital environment, and interacts more with digital content. Such users might be more comfortable clicking on ads, especially if they find them relevant.

Simply put, the more time a user spends online or on a specific site, the higher the number of ads they are exposed to. This naturally increases the chances of them clicking on an ad. Users who spend more time online generate more data, which can be used to tailor ads specifically for them, increasing the likelihood of clicks.

Business Simulation & Recommendation

Feature Importance – Logistic Regression with Scaling Data



TOP 2 FEATURE

- "Daily Time Spent on Site" can be a direct indicator of user engagement. Users who spend more time on a site are more likely to be engaged and possibly more receptive to the content, including ads. They have a higher chance of noticing and clicking on ads compared to someone who just briefly visits the site.
- High "Daily Internet Usage" might indicate that a user is tech-savvy, familiar with the digital environment, and interacts more with digital content. Such users might be more comfortable clicking on ads, especially if they find them relevant.

Simply put, the more time a user spends online or on a specific site, the higher the number of ads they are exposed to. This naturally increases the chances of them clicking on an ad. Users who spend more time online generate more data, which can be used to tailor ads specifically for them, increasing the likelihood of clicks.

Business Recommendation

Based on EDA and feature importance:

Users who click on ads:
1. Spend less time on the site
2. Older
3. Lower income
4. Lower daily internet usage

Users who did not click on ads:
1. Spend more time on the site
2. Younger
3. Higher income
4. Higher daily internet usage

Business Recommendations:

1. **Shorten Session Length**: shortens the duration of a user's session by encouraging them to click on ads earlier in their session. This can be done by placing ads in more prominent places or customizing site content to make it more appealing to users.
2. **Optimize Ad Placements**: Ads placed in areas of the site that users visit frequently in an initial session may have a greater chance of being clicked. Further analysis of user behaviour can help in optimizing ad placement. Ad Optimization for Users with High Internet Usage.
3. **Target Demographics and Ad Personalization** : Displays ads that are relevant to the needs and interests of users in each specific age group and income bracket. Target Users with Low Internet Usage.
4. **Increase Engagement Among Young People**: Display ads specifically designed to attract the interest of young users.

Business Simulation

Rakamin Academy

Before Machine Learning

Goal setting → Ad content creation → Ad placement → Launch campaign → Monitor performance → Calculate ad click ratio

Ad Click Ratio = 50.05%
Profit Percentage = 33.40%

After Machine Learning

Goal setting → Audience segmentation using **Machine Learning** → Ad content creation → Ad placement → Launch campaign → Monitor performance → Calculate ad click ratio

Ad Click Ratio = 98.63%
Profit Percentage = 99.66%

Business Simulation

Rakamin Academy

Cost and Revenue Calculation

assuming the number of users and marketing costs are the same

Cost : Rp 600.000

Revenue : Rp 200.000

Clicked Ads Performance Before Modelling

Total Ad Clicks:	499
Total Target Users:	997
Total Cost:	Rp 199400000.00
Total Revenue from Clicks:	Rp 299400000.00
Total Profit:	Rp 100000000.00
Ad Clicks Percentage:	50.05 %
Profit Percentage:	33.40 %

Clicked Ads Performance After Modelling

Total Ad Clicks:	98334
Total Target Users:	997
Total Cost:	Rp 199400000.00
Total Revenue from Clicks:	Rp 59000400000.00
Total Profit:	Rp 58801000000.00
Ad Clicks Percentage:	98.63 %
Profit Percentage:	99.66 %

Conclusion

- The implementation of the model shows a drastic improvement in the performance of ad campaigns.
- Before modeling, the ad campaign was decently successful, with a conversion of a little over 50%.
- Post modeling, there's a significant rise in the click rate, nearing 99%. The economic gains from this are substantial, with a profit percentage nearing a full 100%.