Predict Clicked Ads Customer Classification By Using Machine Learning

Leading Fintech Presentation

Fadillah Akbar

Data Scientist

Data scientist with a strong mathematical foundation and problem-solving abilities. Experienced in building projects in data mining, data processing, business performance analysis, data visualization, and predictive modeling across multiple industries. Motivated to use data science to improve business impact through analytics, statistics, and machine learning.

www.linkedin.com/in/fadillahakbar

Business Overview

On this occasion we will do machine learning modeling, to predict potential users in digital advertising. We can assume ourselves as a Data Scientist at a company engaged in digital marketing consultant. The business team wants to optimize their advertising methods on digital platforms to get potential users to click a product. So that the costs that will be incurred are not too large. Create a machine learning model that can detect potential users to convert or be interested in an ad. So that we can optimize costs in advertising on digital platforms.

www.linkedin.com/in/fadillahakbar

Table of Content

Customer Type and Behavior Analysis on Advertisement

Data Cleaning & Preprocessing

Data Modeling

Business Recomendation and Simulation

www.linkedin.com/in/fadillahakbar

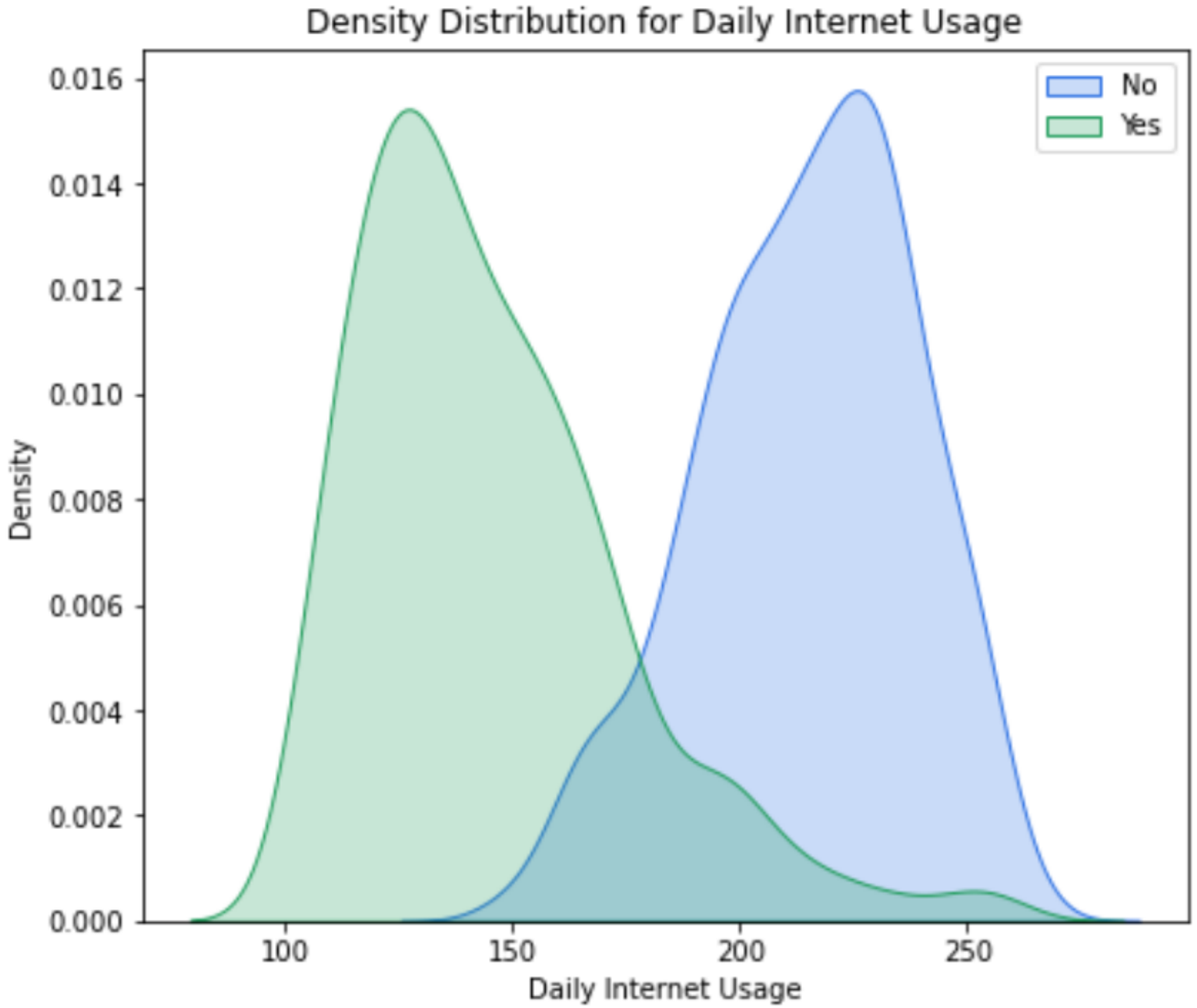
Customer Type and Behavior Analysis on Advertisement

Chapter Description

Analysis of distribution patterns and correlations in order to provide exact information about customer behavior in connection to advertising

Customer Type and Behavior Analysis on Daily Internet Usage

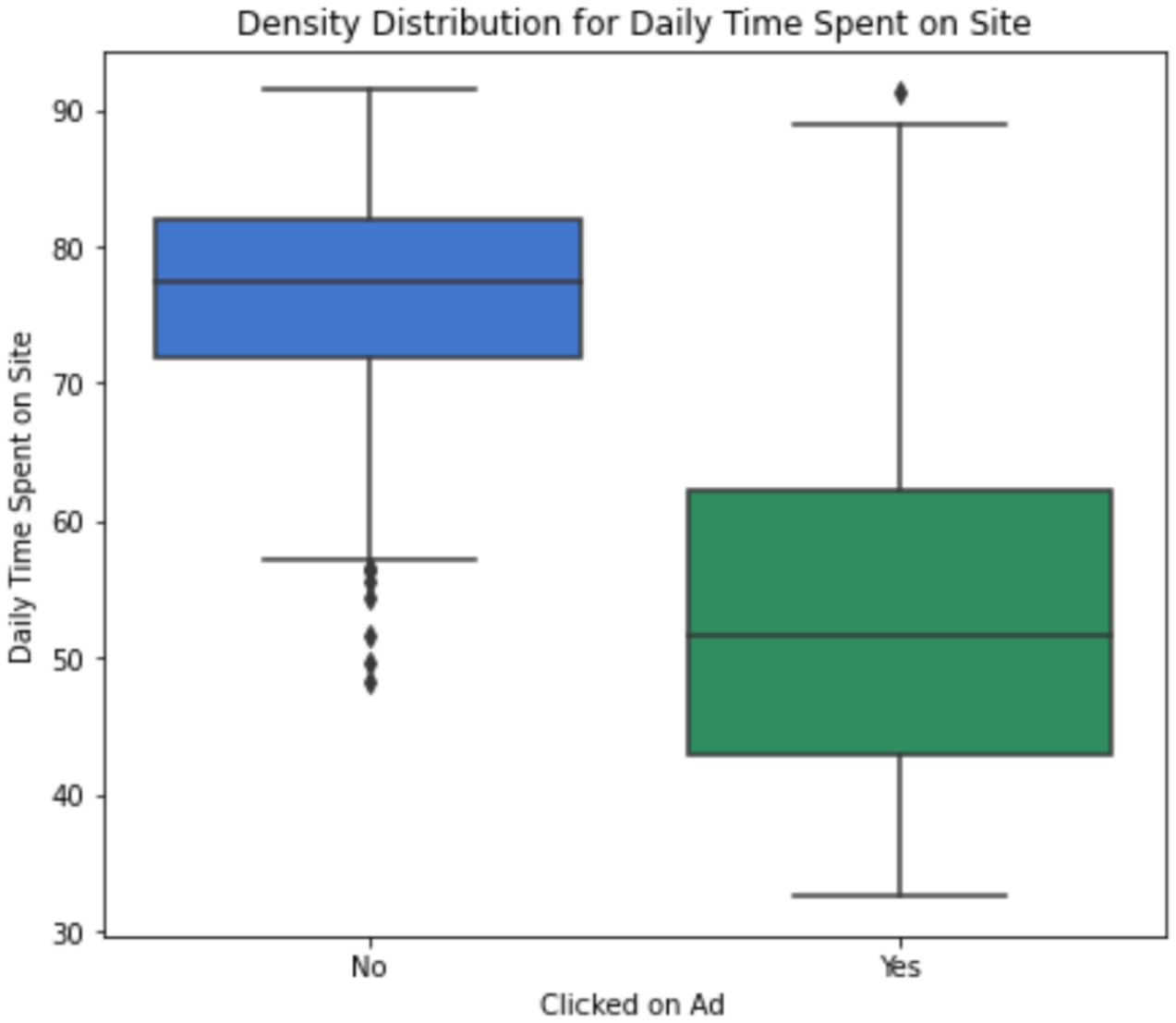
The distribution of daily internet usage may be seen on the side of the EDA (in minutes). The distribution contains several intriguing elements. That users who infrequently use the internet have a higher propensity to click on a product than users who frequently use the internet. This suggests that individuals who infrequently use the internet are more likely to pay attention to adverts on a website.



www.linkedin.com/in/fadillahakbar

Customer Type and Behavior Analysis on Daily Time Spent on Site

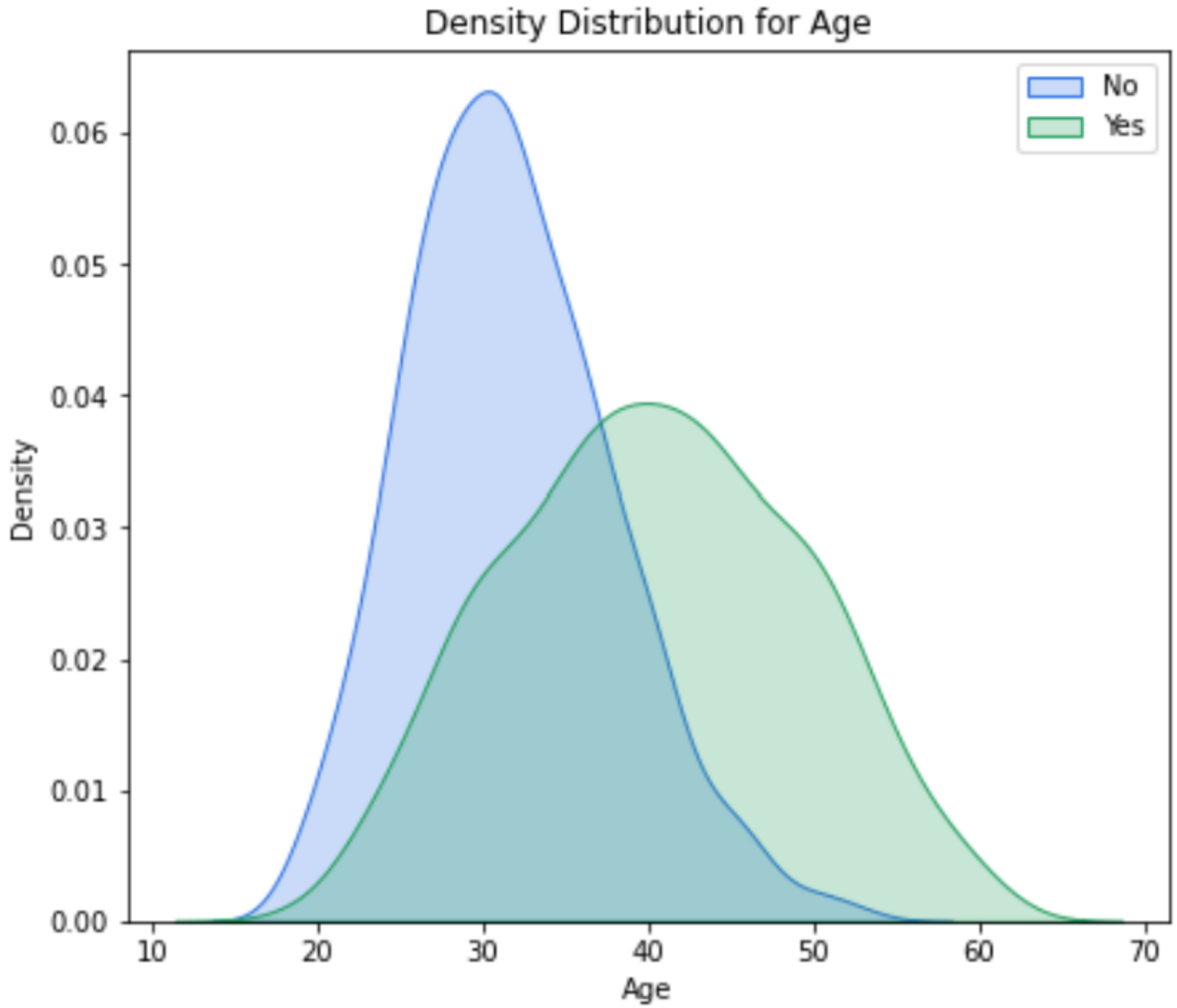
Because internet usage is distributed differently. We want to demonstrate how a user's behavior on a website. According to the EDA on the side, internet consumption and user duration on a website have a similar distribution. That is, potential users can be found even if they only visit a website briefly.



www.linkedin.com/in/fadillahakbar

Customer Type and Behavior Analysis on Age

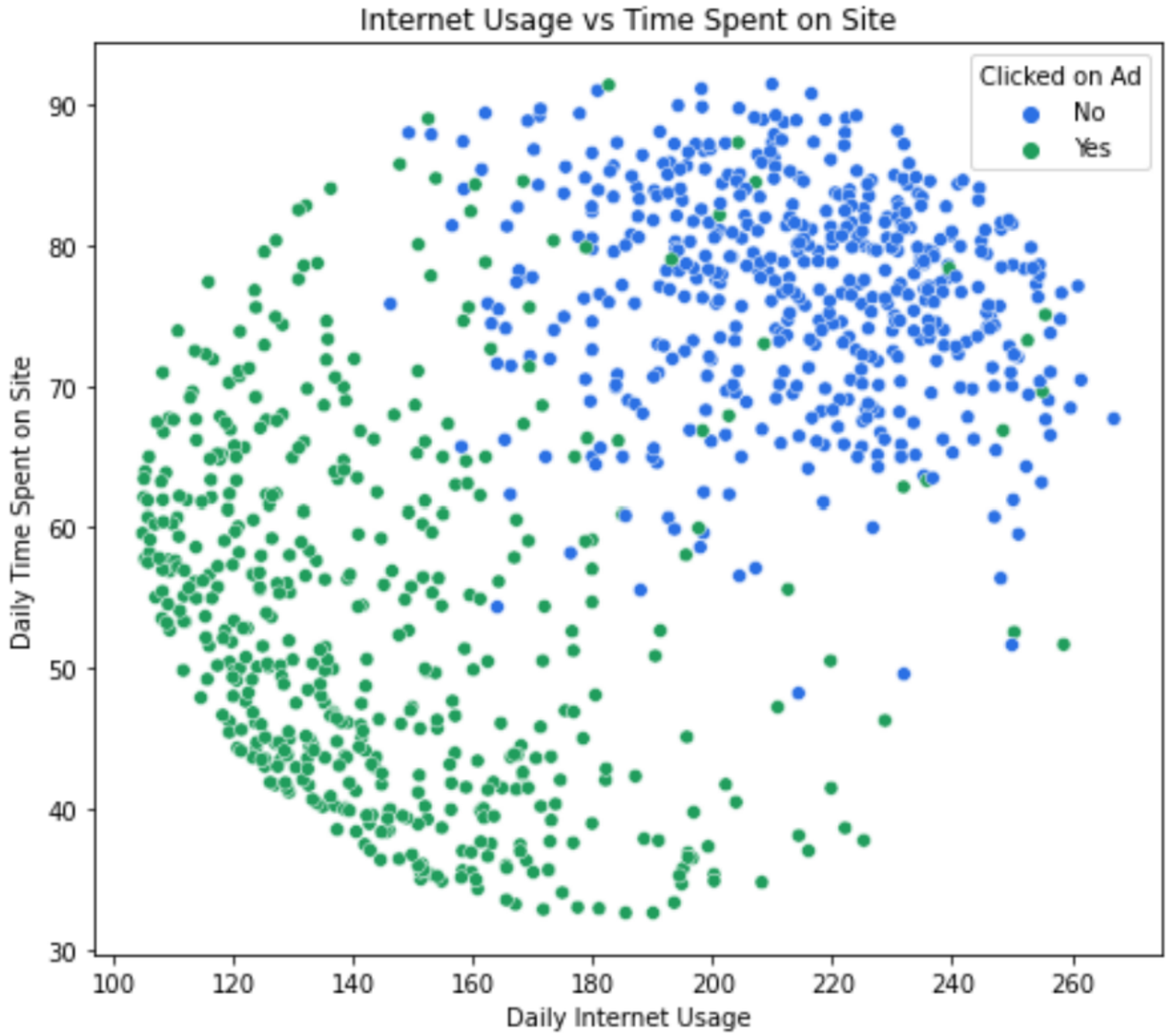
The display of age data reveals that the potential market is really found among older persons. Perhaps because young people are more cautious and picky while using the internet. And young people are quite aware when adverts appear on a website.



www.linkedin.com/in/fadillahakbar

Relationship of Internet Usage and Time Spent on Site

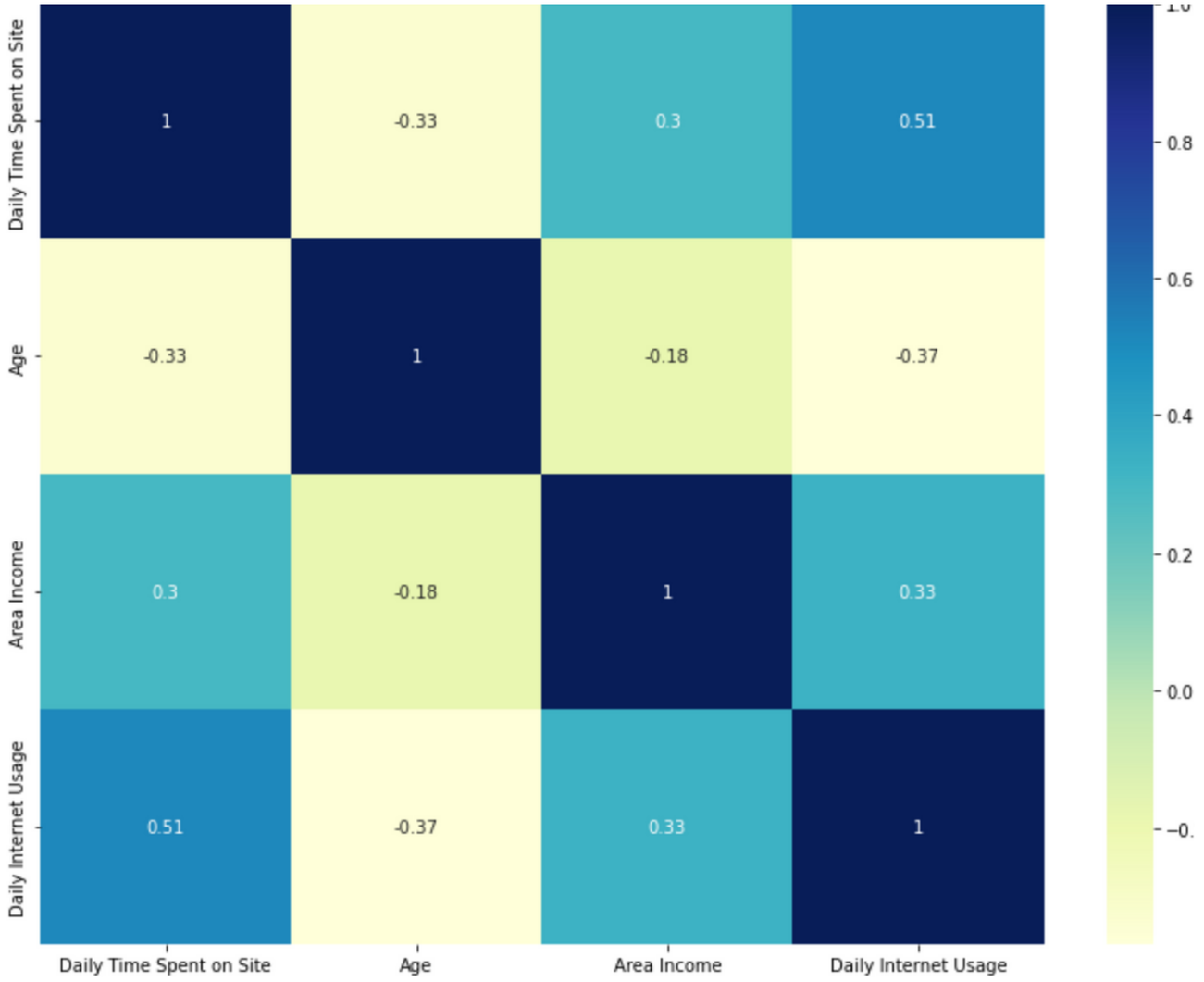
After recognizing the relationship between internet consumption and the length of a website visit. We try to figure out what the connection is between these two traits and their goals. According to the plot, internet usage and the length of a site visit can be classified into two groups: active users and non-active users. These two segments can be a feature that is quite close to whether or not someone will click on an ad. According to the visualization on the side, active users are less likely to click an ad than non-active users. Finally, we can tailor our advertising system to consumers who are not actively utilizing the internet.



www.linkedin.com/in/fadillahakbar

Correlation Between Variables

Because there is no multicorrelation (correlation between variables) in the correlation, we can use all characteristics for modeling. However, we cannot determine the relationship between the feature and the target using Pearson correlation. So, in the following sections, we will utilize PPS (Predictive Power Score) to calculate the link between features and their targets.



www.linkedin.com/in/fadillahakbar

Correlation Using PPS (Predictive Power Score)

Based on the correlation graphic created with PPS, we will just look at the Clicked on Ad feature. Because that variable will be our objective.

The feature is closely related to the target:

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

This correlation plot can be used as a guide for modeling.

www.linkedin.com/in/fadillahakbar

www.linkedin.com/in/fadillahakbar

Data Cleaning and Preprocessing

Chapter Description

Cleaning and processing data into a more useful and efficient format so that it becomes ready-to-use data to facilitate machine learning modeling.

Data Cleaning and Preprocessing

Handle Missing Value

Some features in the dataset contain missing values; one solution is to fill the data value with the average or mode so that the data distribution is not harmed.

Handle Missing Value

```
df['Daily Time Spent on Site'].fillna(df['Daily Time Spent on Site'].mean(),inplace=True)
df['Area Income'].fillna(df['Area Income'].mean(),inplace=True)
df['Daily Internet Usage'].fillna(df['Daily Internet Usage'].mean(),inplace=True)
df['Male'].fillna(df['Male'].mode()[0],inplace=True)
```

Extract Datetime Data

```
df['day_of_week'] = df['Timestamp'].apply(extract_day_of_week)
df['day_of_month'] = df['Timestamp'].apply(extract_day_of_month)
df['month'] = df['Timestamp'].apply(extract_month)

df = df.drop(labels=['Timestamp'],axis=1)
```

Extract Datetime Data

To add new features to the dataset and alter the data type from datetime to numeric, extract datetime data is used. dataset and change the data type to numeric from datetime.

www.linkedin.com/in/fadillahakbar

Data Cleaning and Preprocessing

Split Target and Features

The target and feature are separated so that machine learning can read independent attribute data as well as target/label data predicted by the machine learning model.

Split Target and Features

```
X = df.drop(labels=['Clicked on Ad'],axis=1)
y = np.where(df['Clicked on Ad']=='No',0,1)
```

Get Dummies for All Categorical Features

```
X_dummy = pd.get_dummies(X)
```

One-hot encoding

This method converts category variables into a set of binary variables (also known as dummy variables). get_dummies() is one of the one-hot encoding methods of pandas.

www.linkedin.com/in/fadillahakbar

www.linkedin.com/in/fadillahakbar

Data Modeling

Chapter Description

Selecting the best algorithm to create a machine learning model, as well as determining the evaluation technique and proving whether the model is correct.

Machine Learning Model

First Experiment

Here are the modeling results using the default data (simple preprocessing). The modeling results show that the decision tree classifier has the highest accuracy. Random forest is another method with good accuracy. The accuracy of numerous alternative models, such as logistic regression and k-nearest neighbor, is not as good.

	model_name	model	accuracy	recall	precision	duration
0	K-Nearest Neighbor	KNeighborsClassifier()	0.696667	0.640000	0.721805	0.002985
1	Logistic Regression	LogisticRegression()	0.500000	0.000000	0.000000	0.005012
2	Decision Tree	DecisionTreeClassifier()	0.943333	0.926667	0.958621	0.004999
3	Random Forest	(DecisionTreeClassifier(max_features='auto', r...	0.940000	0.926667	0.952055	0.132503
4	Gradient Boosting	([DecisionTreeRegressor(criterion='friedman_ms...	0.933333	0.913333	0.951389	0.161927

	model_name	model	accuracy	recall	precision	duration
0	K-Nearest Neighbor	KNeighborsClassifier()	0.800000	0.740000	0.840909	0.001001
1	Logistic Regression	LogisticRegression()	0.940000	0.900000	0.978261	0.005002
2	Decision Tree	DecisionTreeClassifier()	0.936667	0.926667	0.945578	0.003999
3	Random Forest	(DecisionTreeClassifier(max_features='auto', r...	0.940000	0.920000	0.958333	0.128998
4	Gradient Boosting	([DecisionTreeRegressor(criterion='friedman_ms...	0.930000	0.913333	0.944828	0.157259

Second Experiment

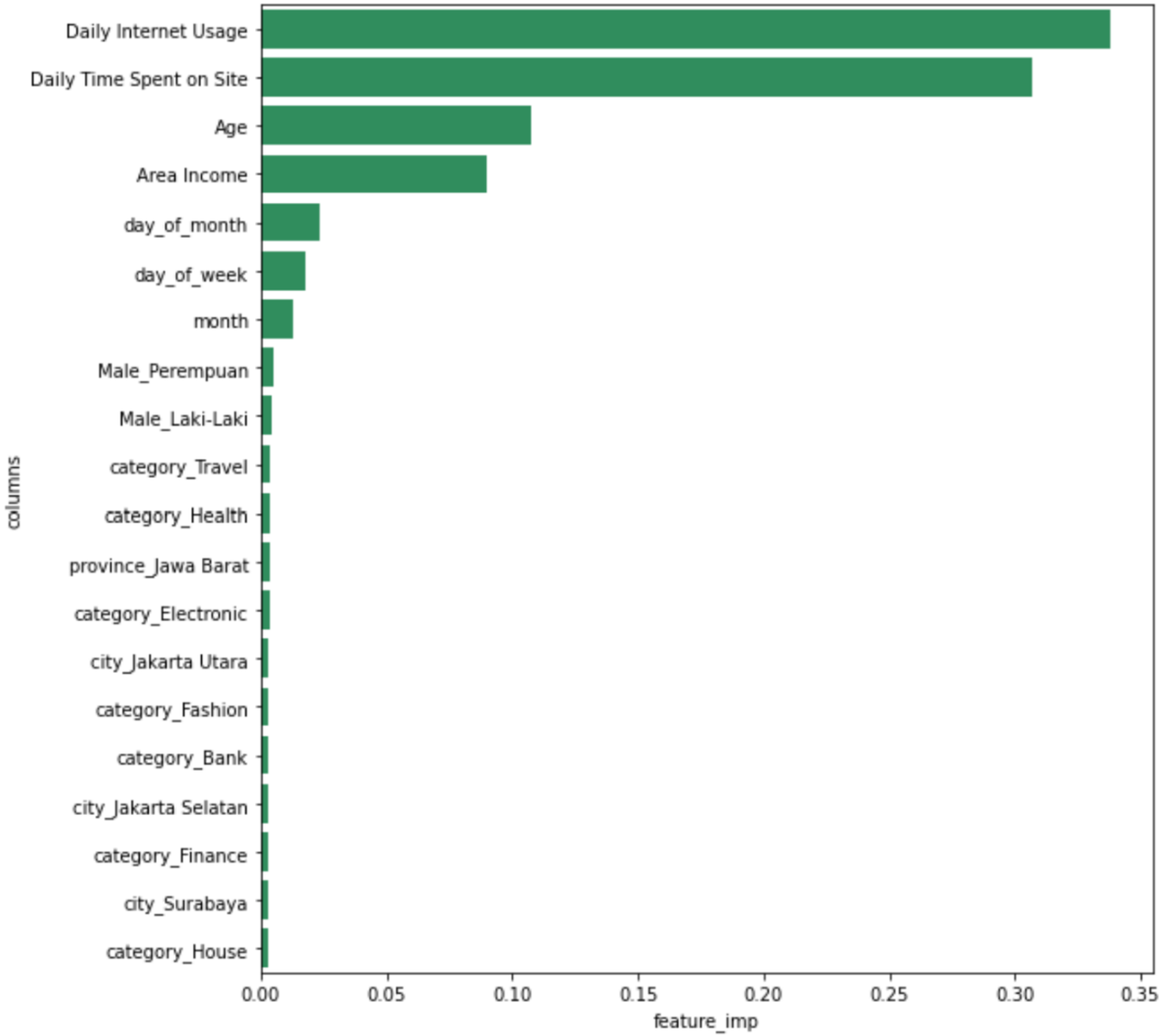
We saw considerable gains in various models after applying data normalization, particularly in the k-nearest neighbor and logistic regression models. After random forest, logistic regression is the most accurate model.We chose random forest as the best model based on these methods because it has the highest accuracy. If there are computing constraints, Logistic Regression is another viable option.

www.linkedin.com/in/fadillahakbar

Feature Importance

Based on the random forest method, we can observe that daily internet usage is a highly crucial factor in deciding whether a person will click or not. Other crucial factors are daily time spent on site, age and income area.

We know from the EDA process that the higher the daily internet usage, the smaller the possibility that the user will click.

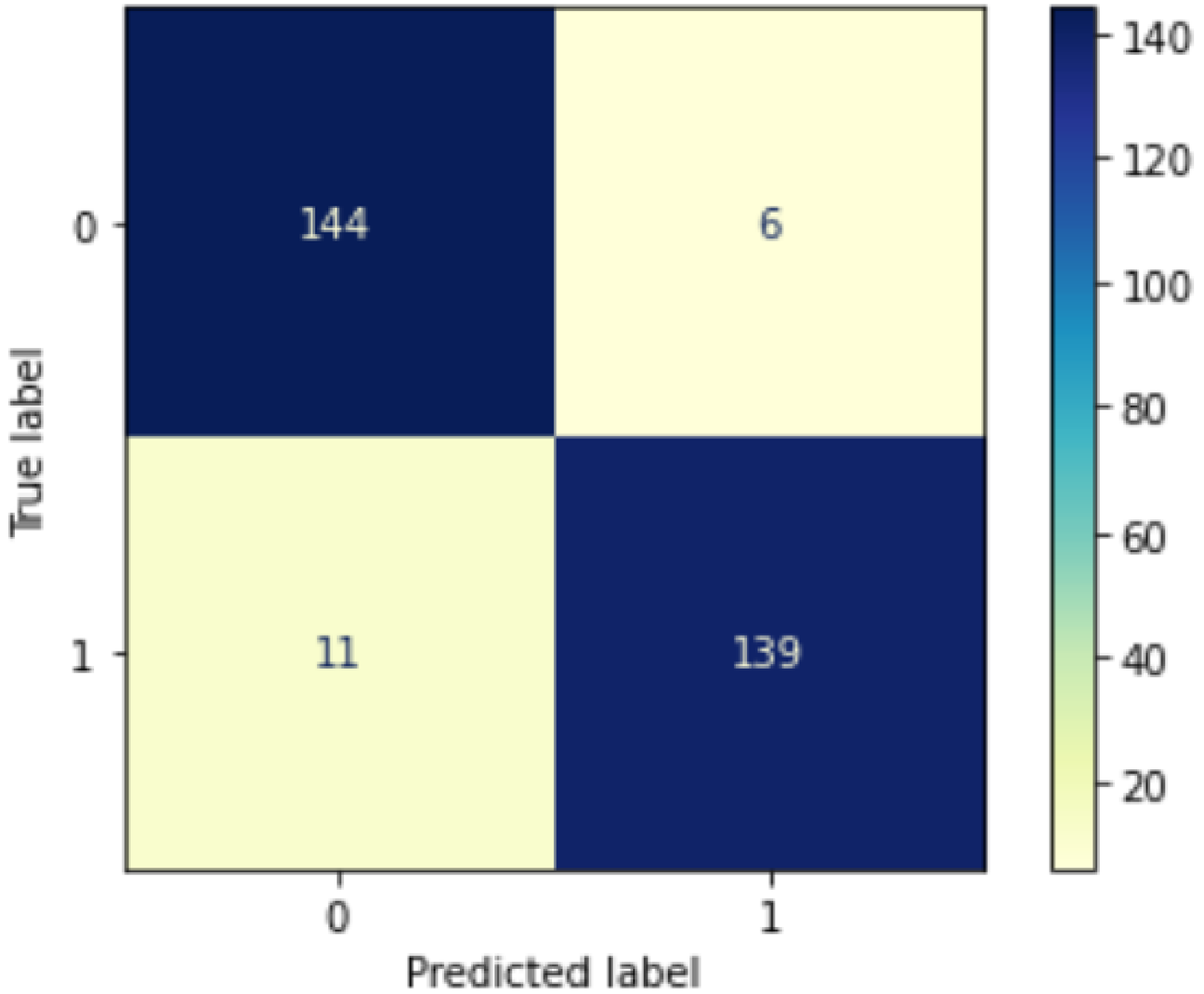


www.linkedin.com/in/fadillahakbar

Model Evaluation

We can see how the random forest model we choose performs in detail by utilizing the confusion matrix. The confusion matrix produced by random forest is excellent.

We can see that there are extremely few prediction errors (purple cells) (top right and bottom left). We will acquire good accuracy, precision, and recall with these findings.



www.linkedin.com/in/fadillahakbar

Business Recomendation and Simulation

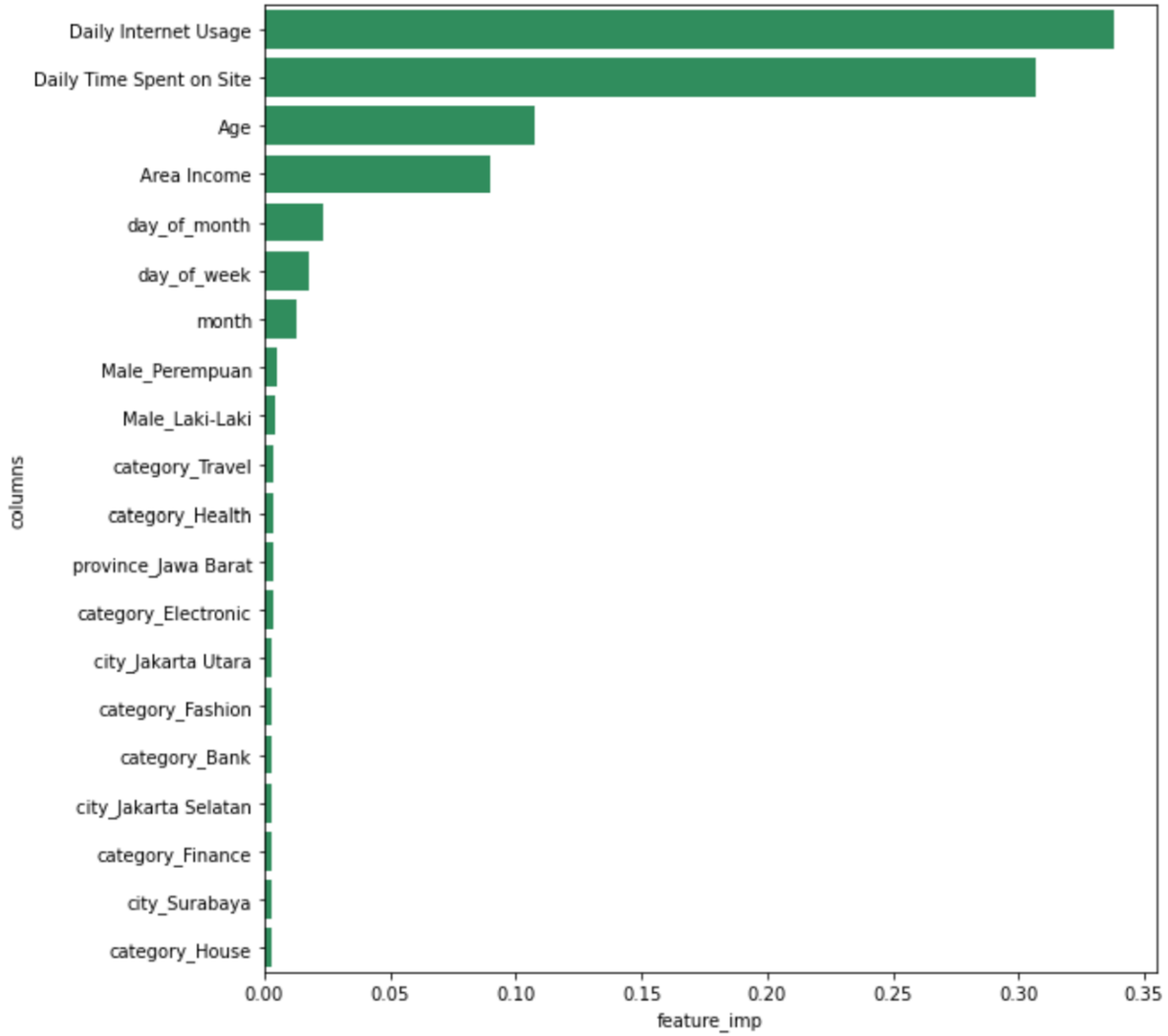
Chapter Description

Develop a simulation and offer a recommendation that has a relationship with marketing

Business Recomendation

We can classify users into two categories based on EDA and feature importance: upper-class and lower-class users.

- The upper class meets the characteristics of regular internet use, frequent visits to a product website, being relatively youthful, and having a high sincome.
- The lower class has the opposite characteristics.



www.linkedin.com/in/fadillahakbar

Business Recomendation

Business Insight

- Users from lower socioeconomic levels are more likely to click on things in digital ads
- Users that use the internet frequently may find it more difficult to receive advertisements because they are accustomed to digital adverts.
- Parents represent a potentially lucrative market for the digital industry.

Business Takeways

- We can use a more unique approach (soft selling) so that users are not as aware of the advertising.
- To attract low-end consumers, employ mainstream material (simple but relevant).

www.linkedin.com/in/fadillahakbar

www.linkedin.com/in/fadillahakbar

Model Simulation

Assumption

To advertise to a user, a budget of 10k rupiah can be used with test data as a simulation tool of around 300 users, with 150 users in each class. We will profit 12k rupiah for every user who converts.

Simulation Without Machine Learning

- We will use a budget of 10,000/user to do advertisement
- Cost : 300 x 10,000 = 3,000,000
- The conversion rate we will get is 50%
- There are 150 users who convert
- Revenue : 150 x 12,000 = 1,800,000
- Profit : 1,800,000 - 3,000,000 = -1,200,000

According to the simulation above, if we do not apply the machine learning model, we will get a potential loss of Rp 1,200,000.-

```
# Ukuran data test
X_test.shape
```

```
(300, 65)
```

```
# Jumlah kelas pada setiap data
jum_class = pd.Series(y_test).value_counts()
jum_class
```

```
1    150
0    150
dtype: int64
```

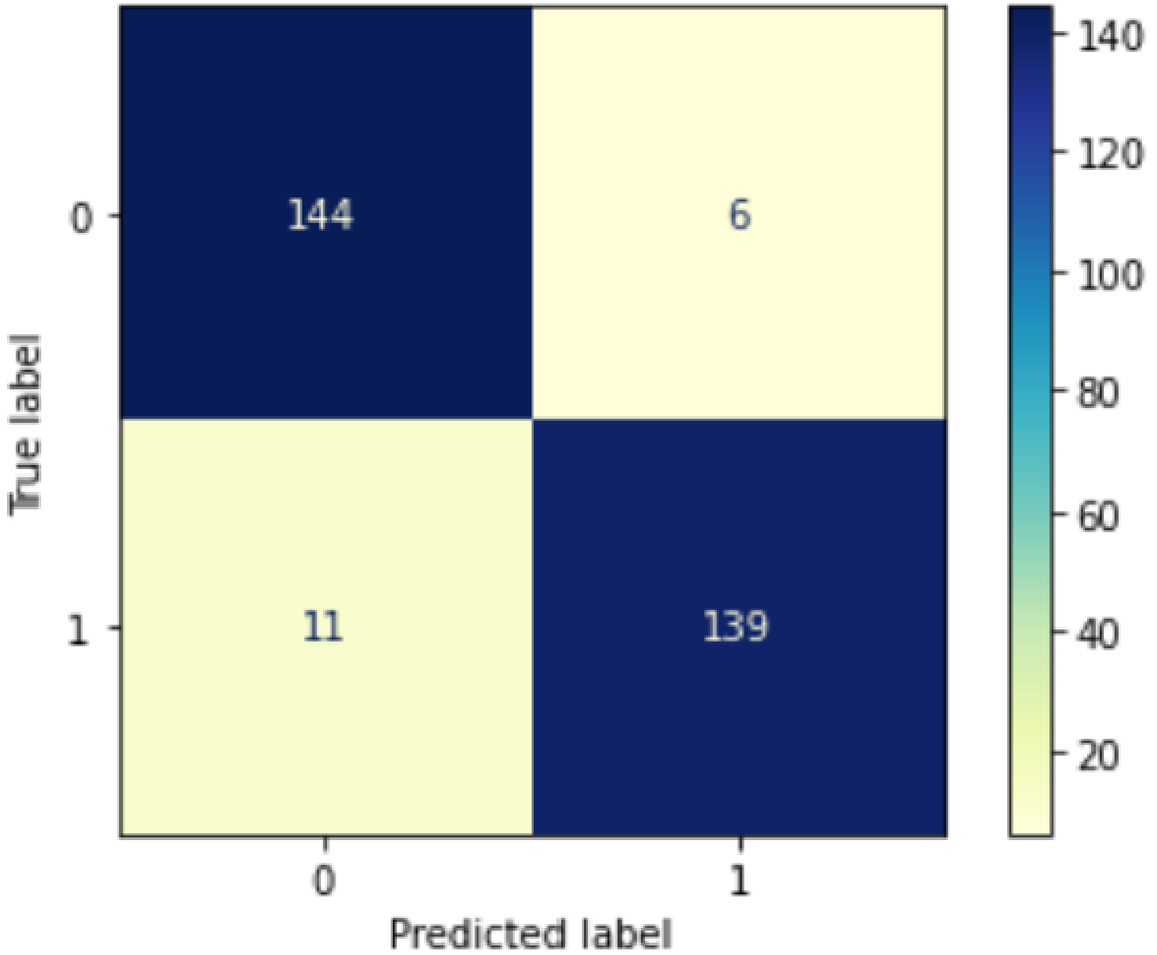
www.linkedin.com/in/fadillahakbar

Simulation With Machine Learning

- We will only advertise to those who have potentially clicked (which we predict 1)
- We will use the same budget which is around 10,000/user to do advertisement
- Cost : 145 x 10,000 = 1,450.000
- We will get a conversion rate of 139/145 = 95.86%.
- We expect that 139 of the 145 users will convert.
- Revenue : 138 x 12,000 = 1,668,000
- Profit : 1,668,00 - 1,420,00 = 218,000

According to the simulation above, if we utilize the machine learning model, we may earn Rp 218,000.-

As a result, machine learning may work more efficiently and effectively, even converting potential losses into possible revenues.



www.linkedin.com/in/fadillahakbar

Thank You

For more details, you can check out the jupyter notebook here.

The dataset can be seen here.

www.linkedin.com/in/fadillahakbar