Ad Click Prediction

By Cristina Iacob

Goal

Explore the relationships between the predicting variables and label

Can we predict who is more likely to click an ad, based on the customer profile and ad headline?

Intended audience

Marketing/Sale teams
Product owners

Data source: https://www.kaggle.com/shubhamsarafo/advertising

Data Description

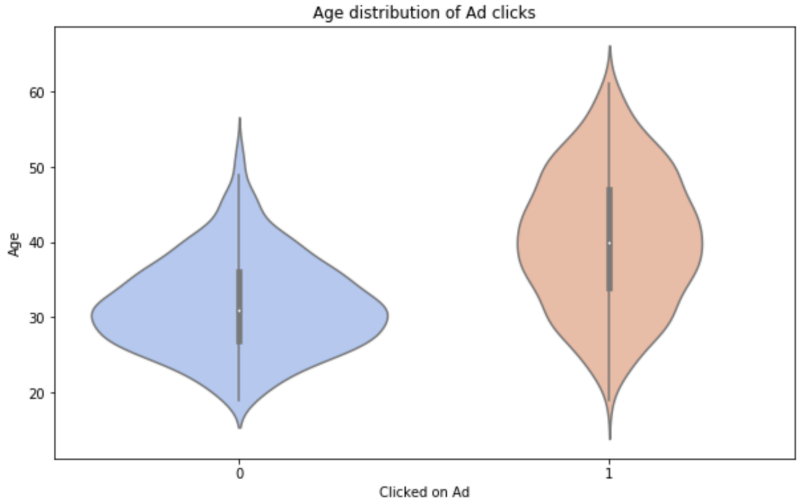
Data set consists of 1000 observations and 10 features:

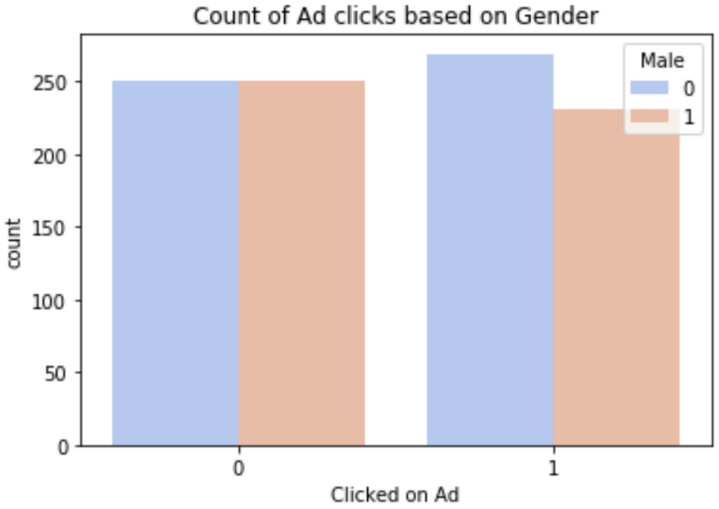
Daily Time Spent on Site	Amount of time in the website by the customer (in minutes)
Age	Age of the customer
Area Income	Avg. Income of geographical area of consumer
Daily Internet Usage	Daily average time spent on the internet by the customer (in minutes)
Ad Topic Line	Headline of the advertisement
City	City of the customer
Male	Indicates whether the customer is male or not (1 - Male; 0 - Female)
Country	Country of the customer
Timestamp	Time at which customer clicked on Ad or closed window
Clicked on Ad	0 or 1 (1 indicated clicking on Ad)

Data source: https://www.kaggle.com/shubhamsarafo/advertising

Visualizing relations between features and label

Which customers clicked more on an Ad?





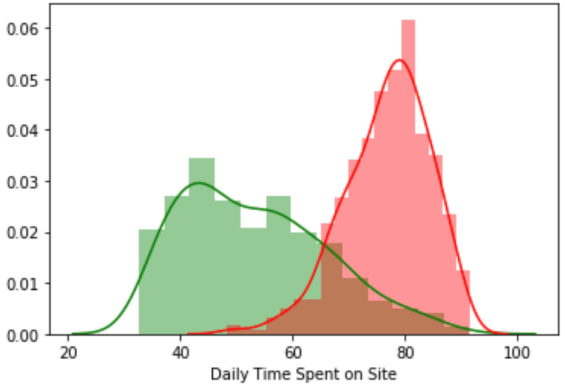
Customers with an average around 40 years old are the most contributors to ad clicking.

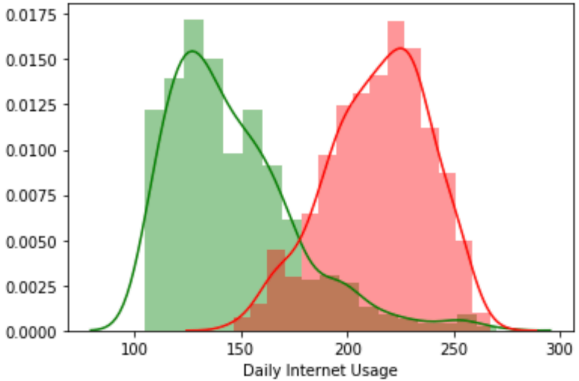
Females tend to click more on Ads.

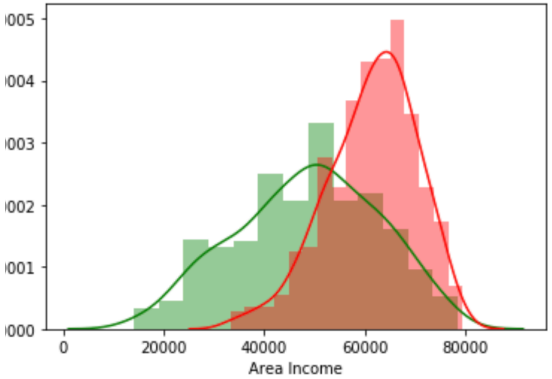
Data source: https://www.kaggle.com/shubhamsarafo/advertising

Visualizing relations between features and label (cont'd)

Which customers clicked more on an Ad?







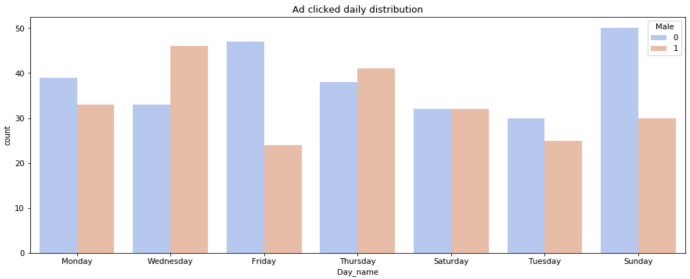
Customers who spent less time on the site and/or browsing the internet

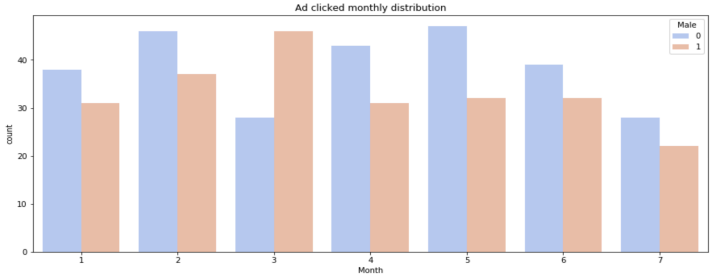
Customers from regions with a lower average income

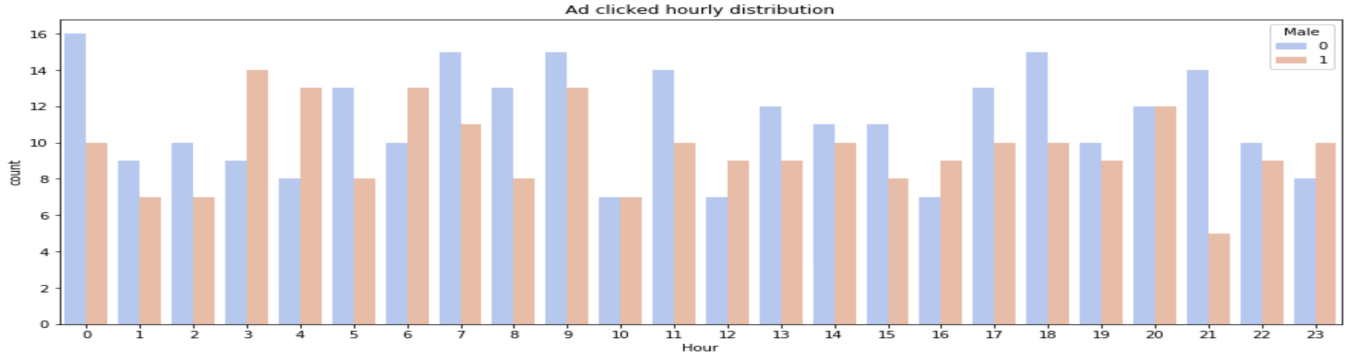
Data source: https://www.kaggle.com/shubhamsarafo/advertising

Visualizing relations between features and label (cont'd)

On which months /days/ hours customers clicked more on an Ad?







Data is only for January to July, 2016.

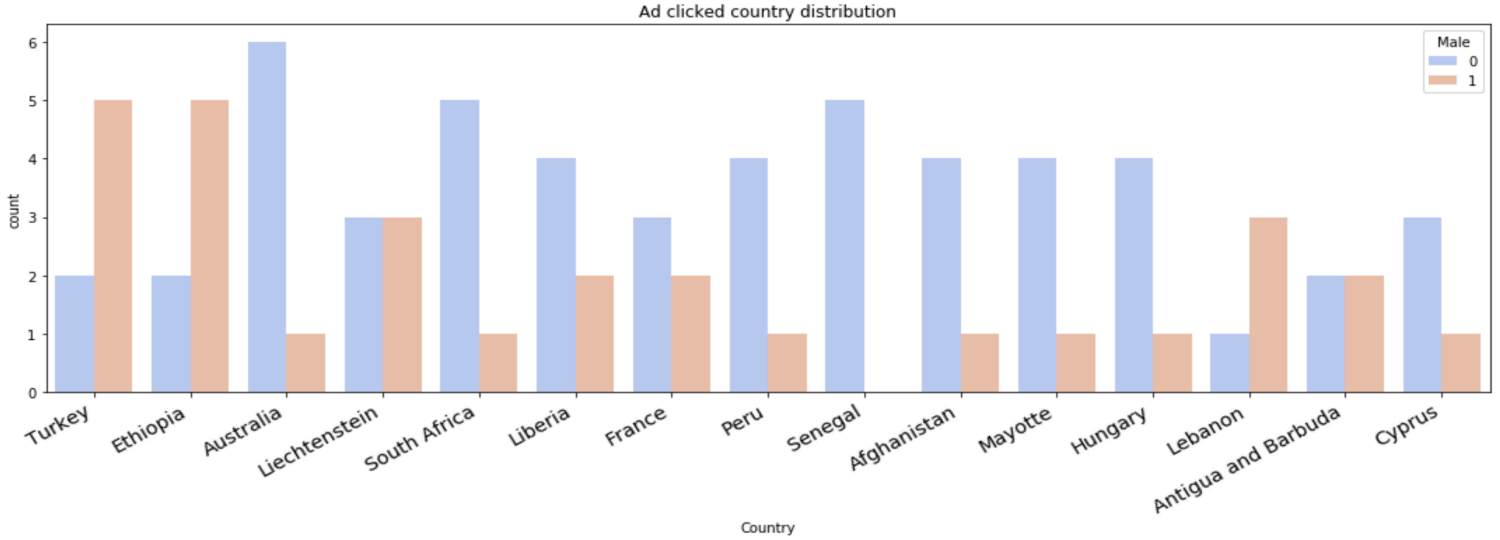
Uneven hourly/daily/monthly frequency, no realistic patterns however

Females are the main contributor on ad click except for March, Wednesdays and Thursdays

Data source: https://www.kaggle.com/shubhamsarafo/advertising

Visualizing relations between features and label (cont'd)

From which countries are customers clicking more on an Ad?



Customers are from 237 countries and 969 cities

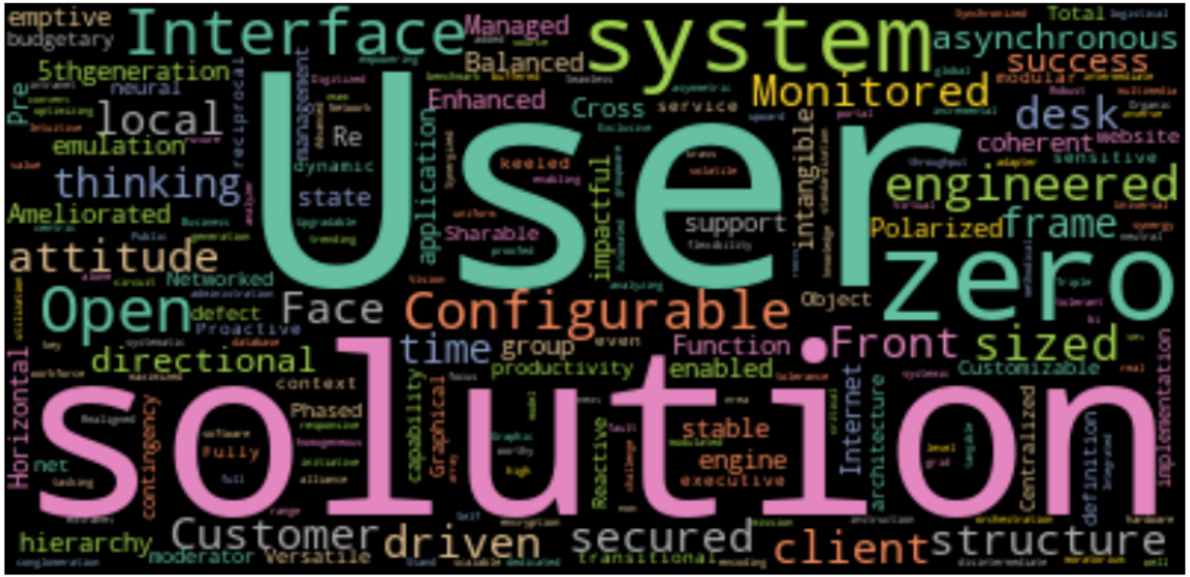
Most developing countries and females are the active contributors.

Data source: https://www.kaggle.com/shubhamsarafo/advertising

What about Ad headlines?

For the 1000 records there are 1000 unique Ad headlines

What about specific words from the Ad title?



Data source: https://www.kaggle.com/shubhamsarafo/advertising

Pre-processing & Modeling

Data:

- Randomly split : 80% for Training and 20% for Testing
- Numerical variables were scaled
- Categorical variables were encoded

Selected Models:

- Logistic Regression
- K Nearest Neighbors Classifier
- Random Forest Classifier
- Gradient Boosting Classifier

Jupyter notebook: https://github.com/cristina-iacob/Predicting-Ad-Click/blob/main/AdClickPrediction.ipynb

Results

	pred_Not_click	pred_click
actual_Not_click	99	1
actual_click	3	97

Logistic Regression

	pred_Not_click	pred_click
actual_Not_click	100	0
actual_click	5	95

K Nearest Neighbors Classifier

	pred_Not_click	pred_click
actual_Not_click	96	4
actual_click	3	97

Random Forest Classifier

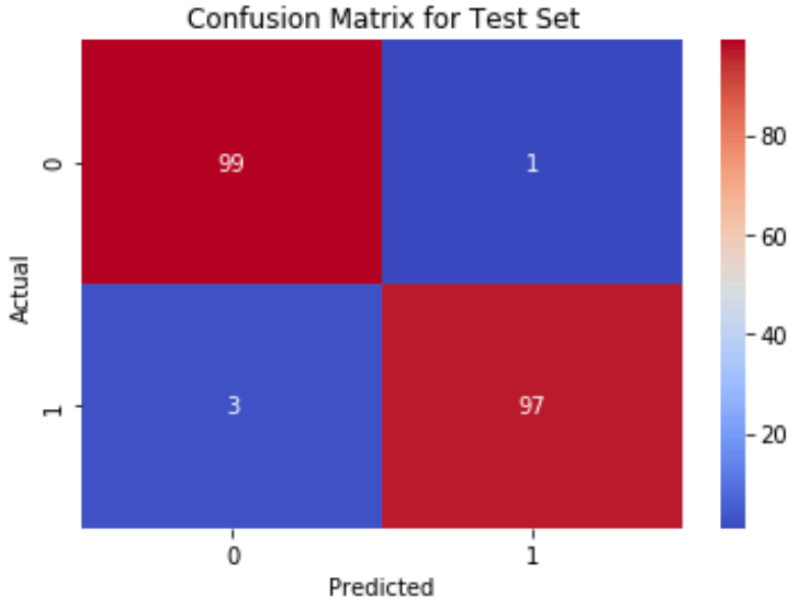
	pred_Not_click	pred_click
actual_Not_click	97	3
actual_click	3	97

Gradient Boosting Classifier

Jupyter notebook: https://github.com/cristina-iacob/Predicting-Ad-Click/blob/main/AdClickPrediction.ipynb

Results (cont'd)

Winner: Logistic Regression



Jupyter notebook: https://github.com/cristina-iacob/Predicting-Ad-Click/blob/main/AdClickPrediction.ipynb

Conclusions:

- According to our analysis, customers who are more likely to click on an ad:
    1. tend to have lower-medium income, between $40.000-$50.000
    2. over 40 years-old
    3. not spending too much time on the website or browsing the internet

- Our model is able to predict with 97% precision the people who will click on the ad and 99% precision those who will not. This means that the model is a bit better to correctly predict people who will not click on an ad then those who will do.

Jupyter notebook: https://github.com/cristina-iacob/Predicting-Ad-Click/blob/main/AdClickPrediction.ipynb

Limitations and Next Steps

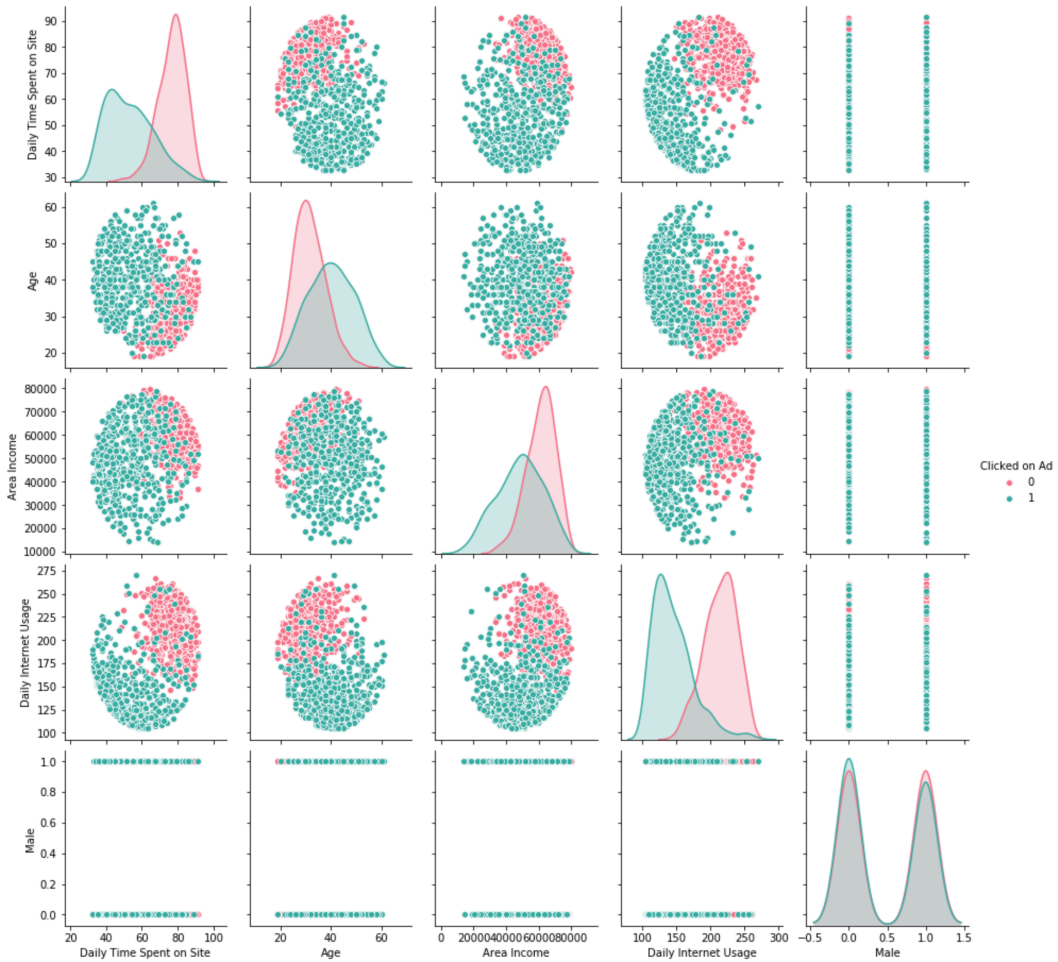
- We had a limited amount of data: 1000 records, half of year
- Ad titles were all unique (and non-sense) so they were not used in modeling; same for Countries and Cities

- New data to include more information such as Education, Income, Marital Status, Ad categories
- Have more testing data and adjust the models based on the new results

Jupyter notebook: https://github.com/cristina-iacob/Predicting-Ad-Click/blob/main/AdClickPrediction.ipynb

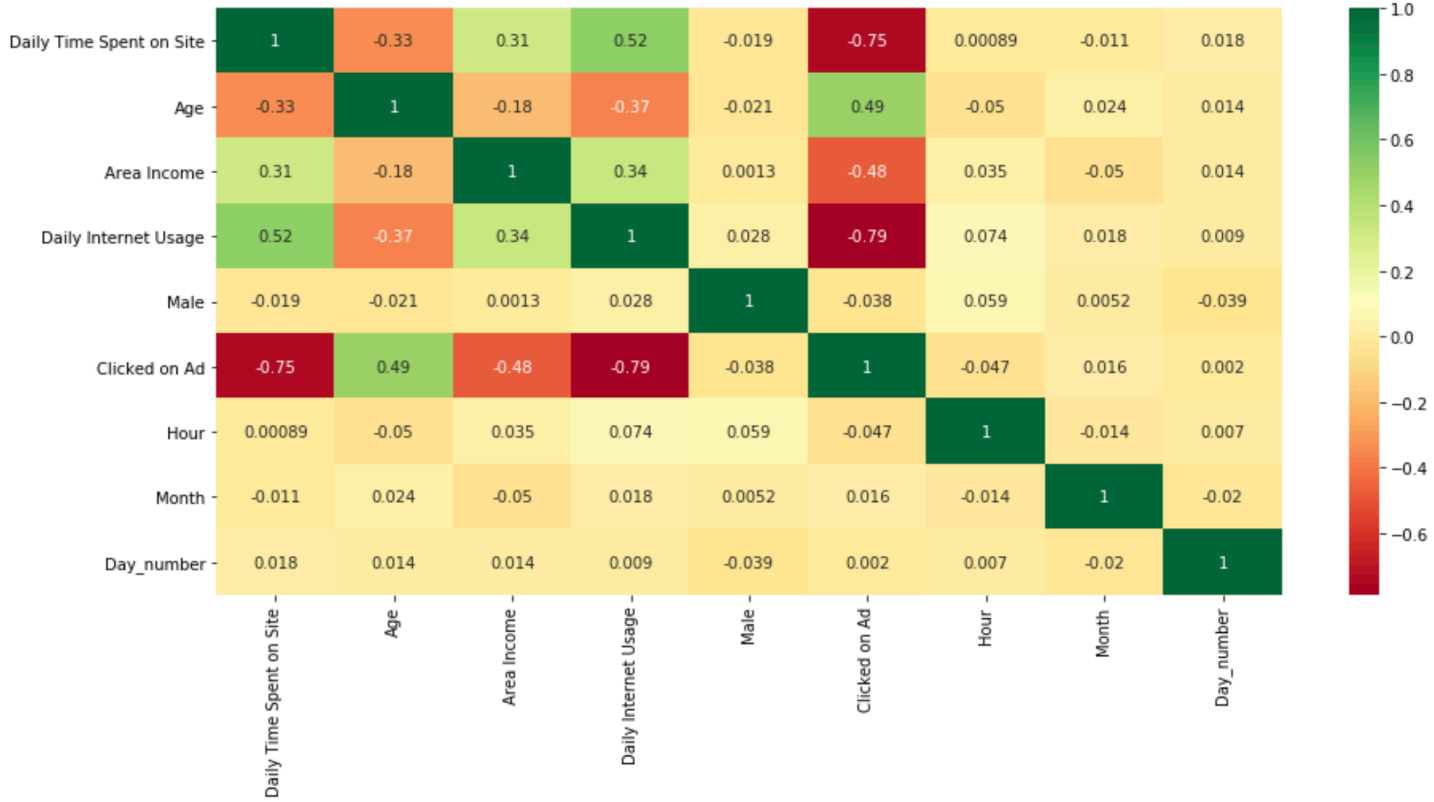
Questions?

Additional Resources



Numerical variables visualization

Correlation matrix



Jupyter notebook: https://github.com/cristina-iacob/Predicting-Ad-Click/blob/main/AdClickPrediction.ipynb

Visualizing our mistakes

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Male	Hour	Month	Day_name	clicked	click_pred	error
702	87.27	30	51824.01	204.27	1	6	3	Sunday	1	0	True
305	79.81	24	56194.56	178.85	1	11	5	Tuesday	1	0	True
181	84.53	27	40763.13	168.34	0	21	1	Monday	1	0	True
998	55.55	19	41920.79	187.95	0	2	3	Thursday	0	1	True

```
print(ad_df.loc[[998]])
```

```
     Daily Time Spent on Site  Age  Area Income  Daily Internet Usage  \
998                     55.55   19     41920.79                187.95

                          Ad Topic Line         City  Male    Country  \
998  Proactive bandwidth-monitored policy  West Steven     0  Guatemala

     Clicked on Ad  Hour  Month  Day_number  Day_name
998              0     2      3           3  Thursday
```

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Male	Clicked on Ad
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000
mean	65.000200	36.009000	55000.000080	180.000100	0.481000	0.50000

Jupyter notebook: https://github.com/cristina-iacob/Predicting-Ad-Click/blob/main/AdClickPrediction.ipynb

Logistic Regression Classification Report

	precision	recall	f1-score	support
0	0.97	0.99	0.98	100
1	0.99	0.97	0.98	100
accuracy			0.98	200
macro avg	0.98	0.98	0.98	200
weighted avg	0.98	0.98	0.98	200

- Precision – outcomes correctly predicted: higher for predicting “ckick”
- Recall – actual positives correctly identified: higher for predicting “no click”
- F1 – score – weighted average of precision and recall: same value for predicting “click” and “no click “

Jupyter notebook: https://github.com/cristina-iacob/Predicting-Ad-Click/blob/main/AdClickPrediction.ipynb