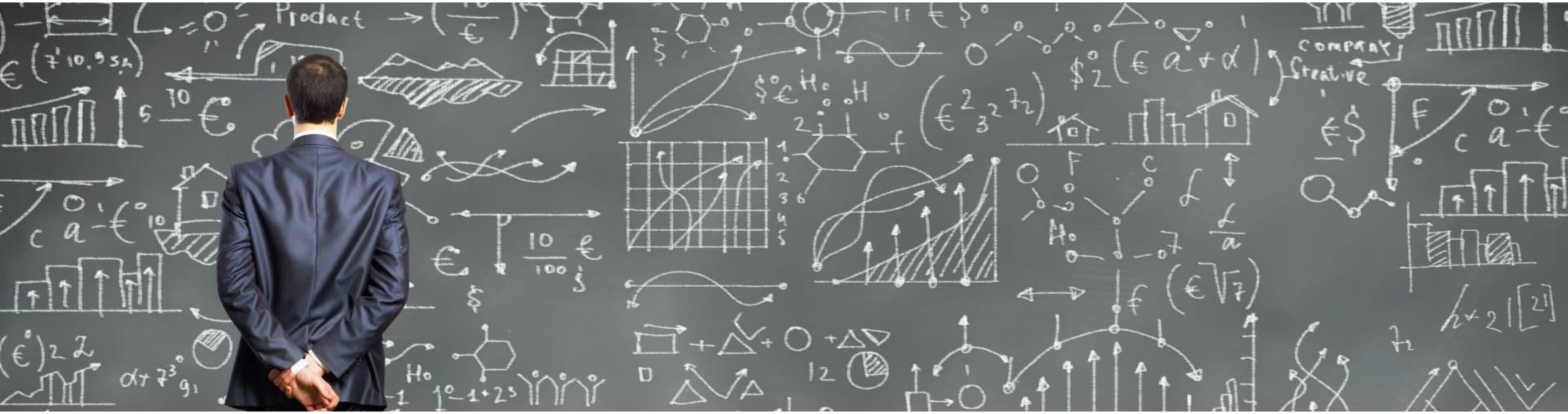
Hellen Wainaina



Machine Learning Classification

Predict the likelihood that a consumer clicks on an ad

Agenda

Improving Audience Targeting

Background / Business Problem	3
Executive Summary / Key Takeaways	4
Analysis – Data Set Characteristics	5
Analysis – EDA	6-11
Analysis – Cleaning & Pre-processing	12,13
Analysis – Modelling, Tuning & Evaluation	14
Analysis – Key Results & Recommendation	15
Next Steps & Improvements	16
Appendix	17,18

Business Problem

Situation

Company A would like to create awareness about product M and is interested in working with Media conglomerate company X. X has multiple websites within its network but only 10 have the right audience and provide the relevant contextual environment for company A.

The goal is to increase awareness of product M and therefore accuracy in correctly predicting whether a consumer will click on the ad is key.

Complication

Of the 10 websites, company A is only interested in placing ads on 2 websites where they will get the highest number of clicks.

If the goal is not met by this campaign company X will not secure the full amount to run the ad since company A is willing to pay $25 per click up to a max of $40,000. They will also potentially lose any future business with the company estimated at a minimum of $300,000 per year.

Executive Summary

- The baseline model has an accuracy of 90% and therefore the model built in this project has improved the accuracy by 7% over the baseline (at 97.25%)

- The key features in predicting the likelihood that a person on the site would click on an ad were time spent online and, on the website. This information is useful in improving audience targeting

- Data with more relevant features would be useful in further increasing accuracy

Data Set Characteristics and Information

CHARACTERISTICS

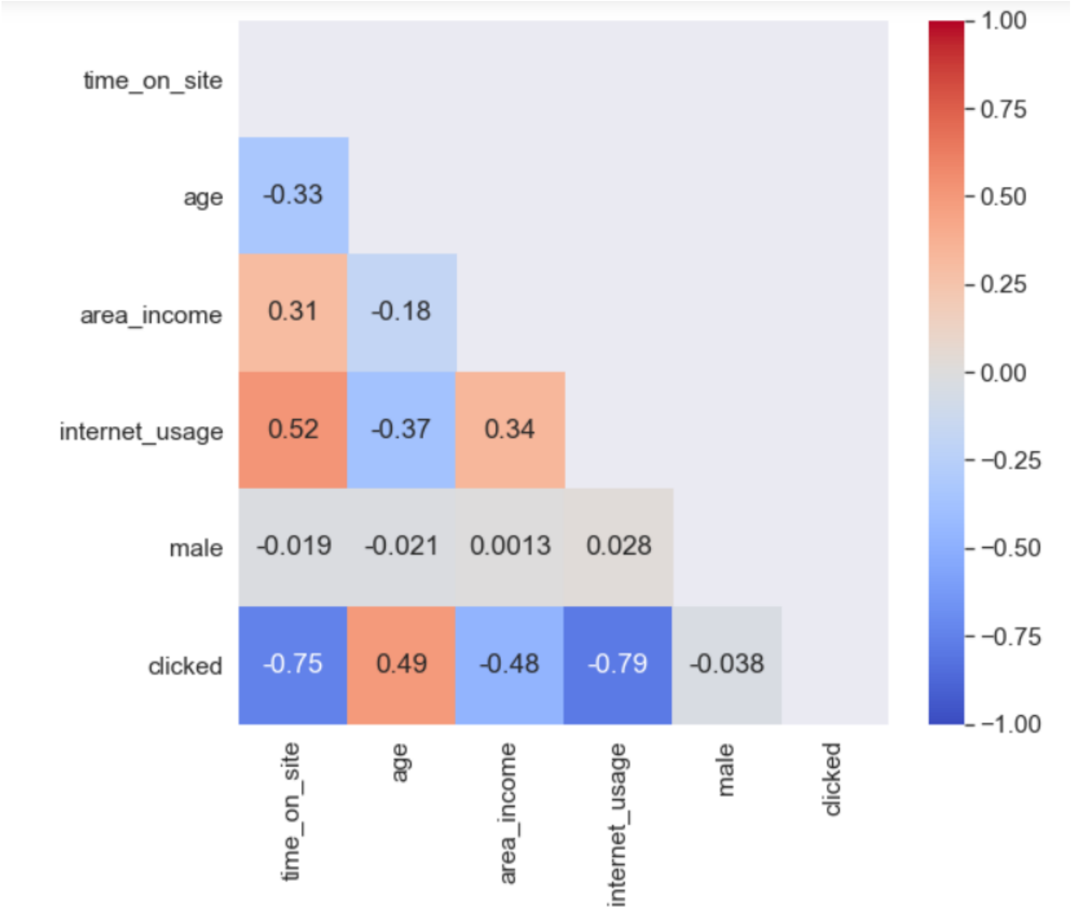
- The dataset used for this project has 1018 rows and 10 columns.
- There were 4 missing values in the target variable column and 7 duplicated values.

FEATURES

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

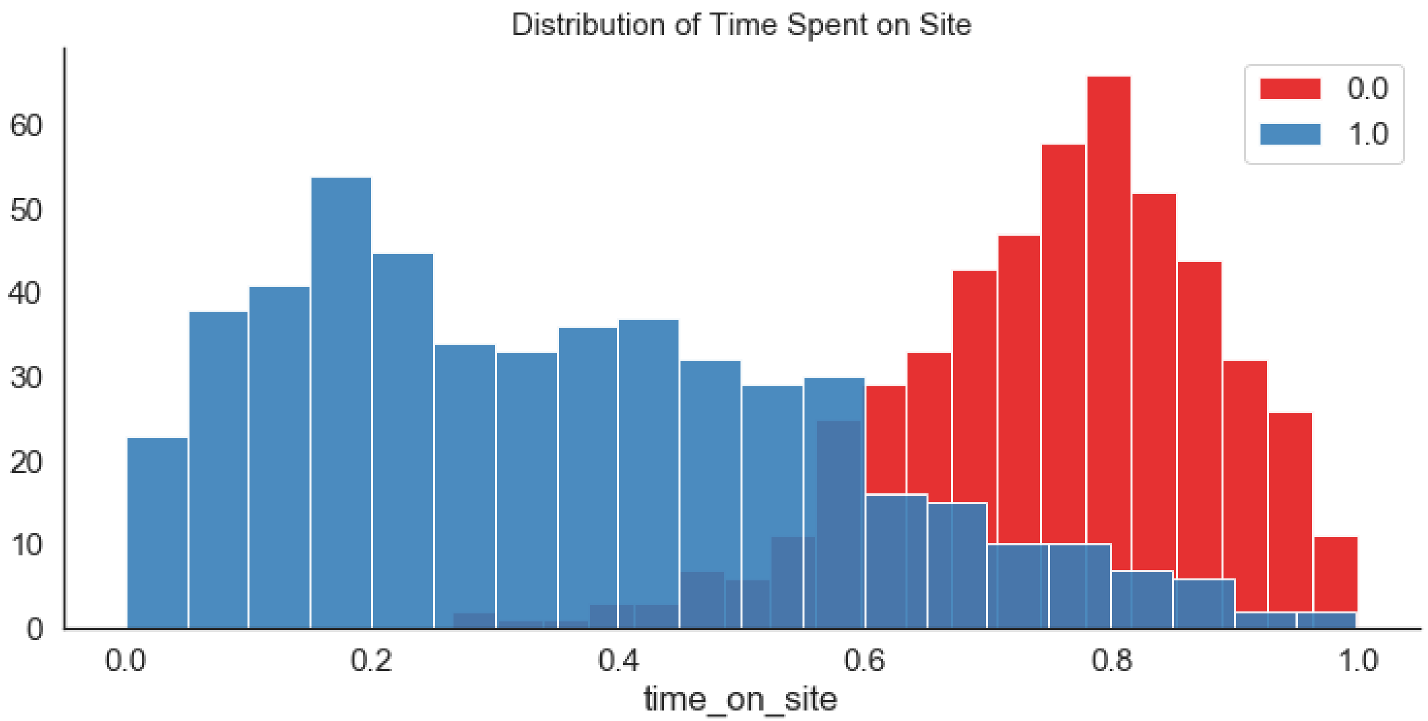
EXPLORATORY DATA ANALYSIS

Internet Usage and Time on Site highly negatively correlated with target variable



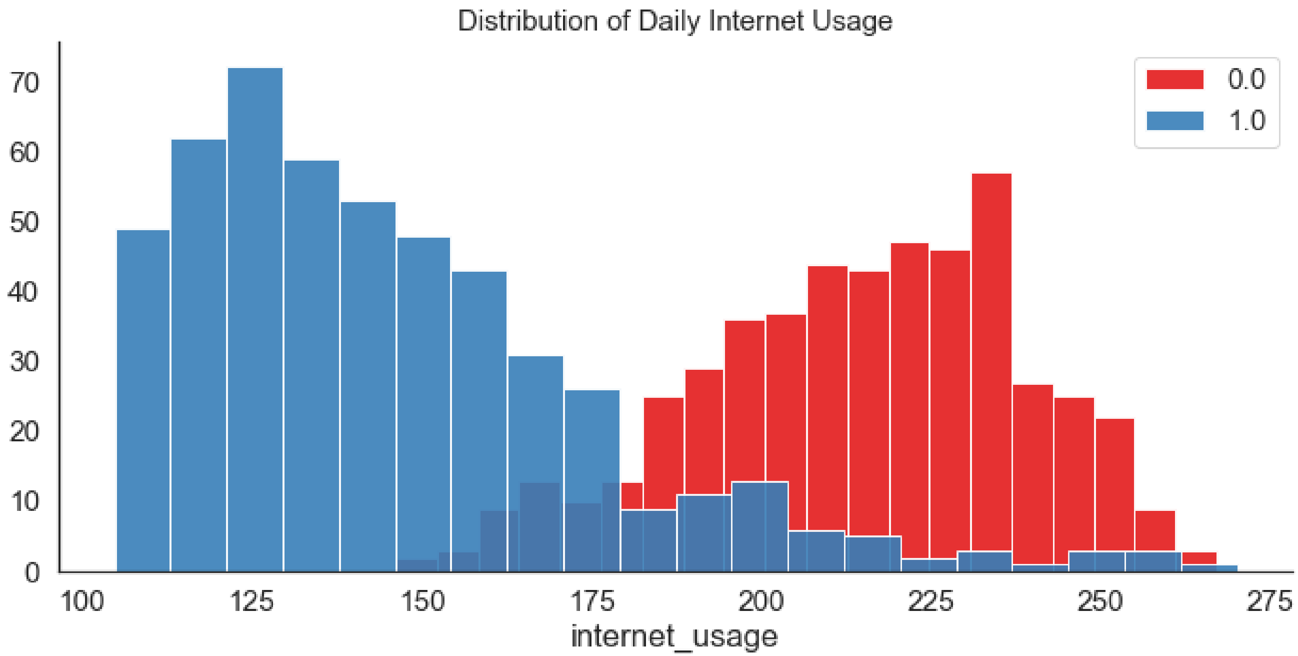
- There is a strong negative correlation between the target variable and other numerical features. Gender shows almost no correlation
- Time on site and internet usage have the highest negative correlation with the target variable
- Age has a positive correlation with the target variable whereas area income has a negative correlation

Consumers who spent less time on site more likely to click on the ad



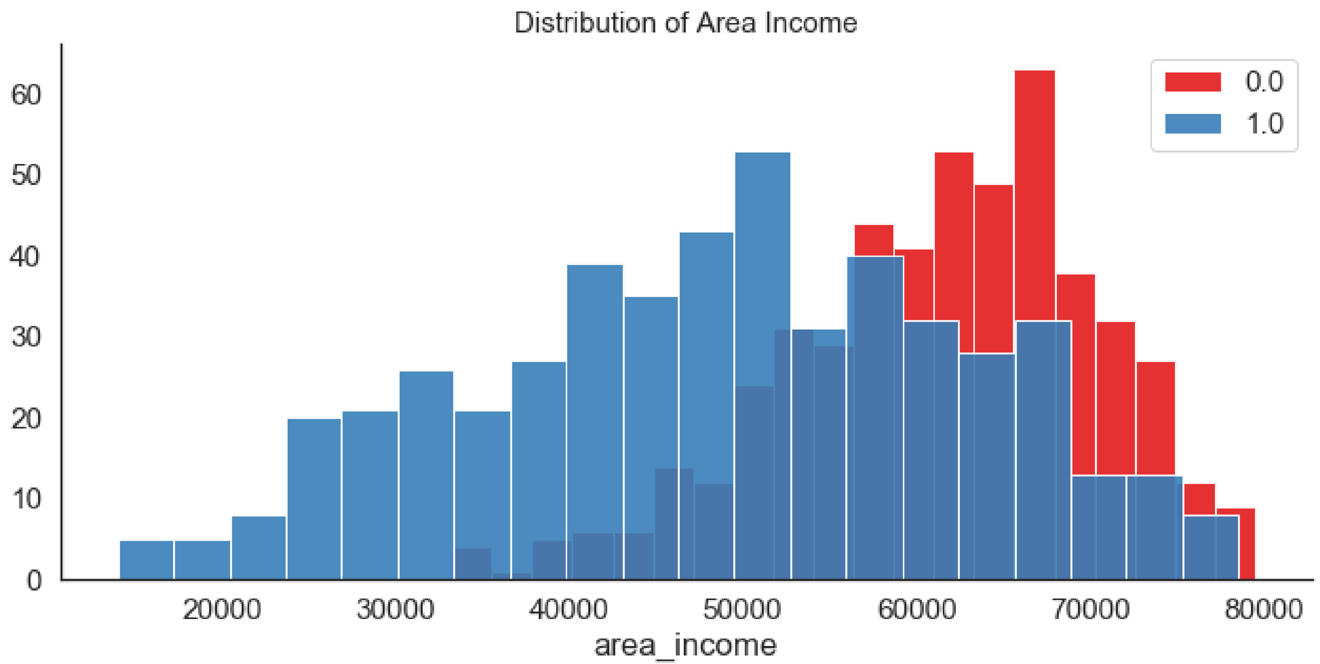
- Time on site is likely to be an important feature in our model since there is a good distinction between the people who clicked on the ad and those that did not
- The above chart further supports that there is a strong negative correlation between time spent on site and likelihood of clicking on an ad. It is important to note that even though the consumers who spent less time clicked on the ad, the minimum amount spent on the site was ~32 mins

Audiences who spent less time on the internet were more likely to click on the ad



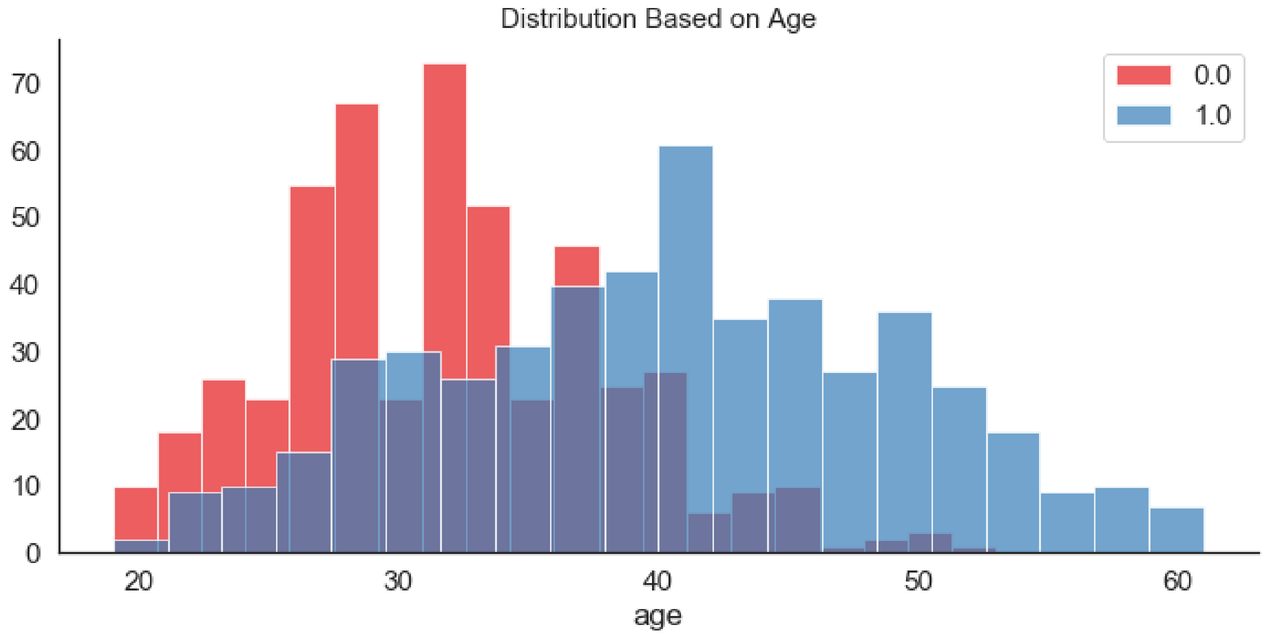
- There's a distinction between consumers who clicked on the ad and those that did not and therefore there's a high likelihood that internet usage will be an important feature in our model
- Consumers who spent less time on the internet were more likely to click on the ad.The minimum amount of time spent on the internet in this case is ~23 minutes

Audiences in high income areas less likely to click on the ad



- There's an even distribution for consumers who are likely to click on the ad. The average area income for people in this dataset is ~$55K.
- Consumers who were less likely to click on the ad live in higher income areas. A likely explanation would be that the consumer living in a higher income area does not resonate with the ad or with the product

Younger audiences were less likely to click on the ad



- There's an even distribution of people who are likely to click on the ad. However, younger consumers were less likely to click on the ad which might be due to the product/service sold by company A
- A brand study might be useful in revealing more information

Data Cleansing & Pre-processing

Categorical Features

- This dataset contains two categorical features, the target variable (whether the audience clicked or did not click) and the gender (whether the audience was male or not).
  - Both are almost even at around 50% each.
- Since the number of observations in the dataset that clicked on the ad is about 50% and we only have two classes, to preprocess the data we do not carry out resampling

- The target column had 4 null value and those were removed from the data set entirely since it would not be accurate to impute values in a target class. However, it would still be helpful to investigate why they were missing in the first place

Data Cleansing & Pre-processing

Numerical Features

- This dataset consists of data on time spent on site, internet usage, area income and age
- Age and area income had skewed distributions and therefore were log transformed to avoid introducing bias to the model

Other Features

- Ad topic line would be best analyzed using a natural language processing model which merits its own project and is not pursued here
- Timestamp would also be best analyzed using a time series model
- The dataset contains too many countries (237) and too many cities (969) and the assumption is that they will not have a very strong predictive power and are thus not used in creating the model

Modelling, Tuning & Evaluation

Model Selection

- Since this is a classification problem, a logistic regression model was used
- In terms of hyper parameter tuning, the initial range focused on a C value between 0.001 and 1000. The results showed that a value of 1 was closer to the best value of C. The next evaluation was with values between 0.01and 10 and the best value was 3.34 which was the value used in this model

Model Evaluation

- The baseline model had a 90% accuracy, the final model used here has a 97% accuracy
- Initially the model had an accuracy of 95.77%, after hyperparameter tuning this increased to 97.25%

Model Performance Results

The model selected had a final accuracy score of 97.25% and an F Score of 98.25%. In future, this model can be tuned further, and additional features if available could be useful. This would likely improve performance

Analysis Results & Recommendations

Key Predictors

- Internet usage and time on site were the most important predictors of whether the consumer would be more likely to click on the ad.The next in order of importance were age and income. Gender had an almost negligible effect on predicting the outcome.

Model Performance

- The model has an accuracy of 97.25% and an F score of 98.25%

Business Recommendation

- With a 97.25% accuracy, the goal of about 1600 clicks should be achieved since most websites make more than 100,000 impressions
- Consider other factors such as contextual environments where the ad is placed and also how viewable the ad is

Next Steps & Improvements

- This dataset was originally posted on Kaggle, however the dataset used in this project is from DSDJ

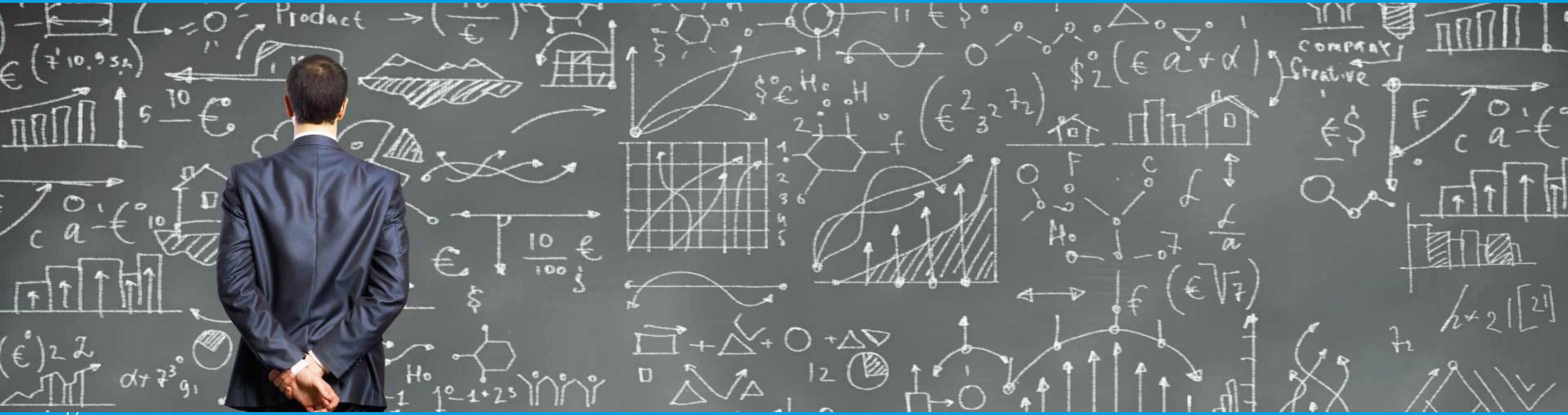
Project/Approach Improvements

1. Consider testing more models such as gradient boosting and random forest to determine their accuracy score
2. A time series model would be useful in determining the time the ad was clicked on the most
3. It would also be helpful to carry out natural language processing to determine whether there was a specific ad topic line that more consumers resonated with

Lessons learned

1. Although time spent online is a useful indicator as to whether someone would click on the ad, this is obvious since one would have to be online to click on the ad. More useful indicators would be the education background of the consumer and their interests and affinities

Appendix



Assumptions

1. The dataset is balanced at 50%.Real world datasets are not likely to be balanced. CTRs are usually very low at about 1-3% of the impressions served and therefore more representative data would be useful
2. The company is compensated based on the number of clicks; most websites are compensated based on the number of impressions made
3. The ad is a homepage takeover since that would be the most viewable
4. Creative concept and size have no impact on the likelihood that an audience will click on the ad