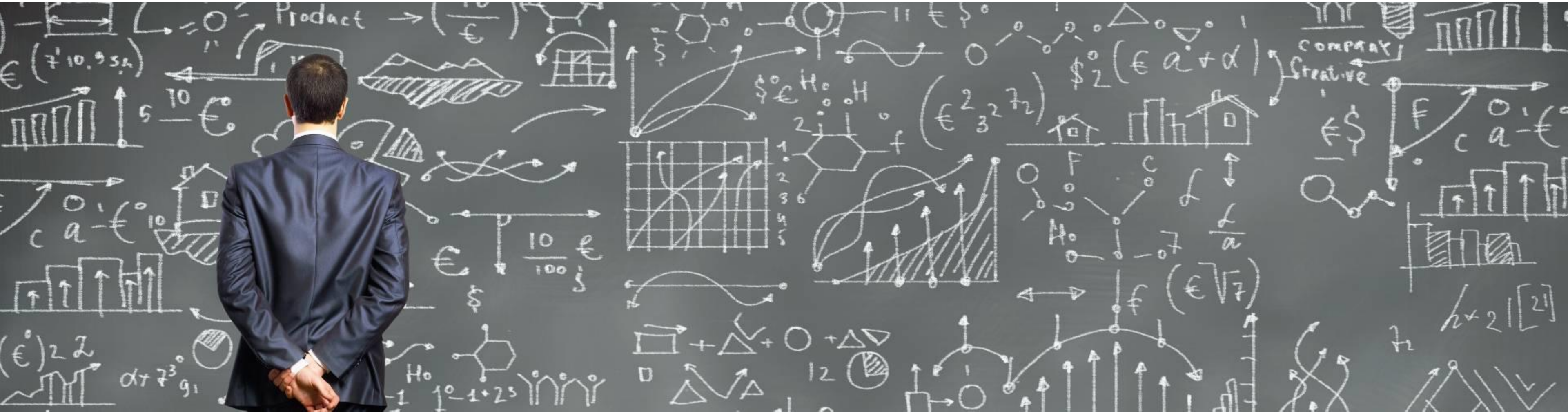
Hellen Wainaina



Machine Learning Classification

Likelihood that a consumer clicks on an ad

Agenda

Improving Click Through Rate

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	17-19

Business Problem

Situation

Media conglomerate company X has a network of different websites within its network and would like to run a certain ad for company A which is selling product M.The company is willing to pay the media company $25 for each successful click up to a maximum of $40,000 and is looking to create awareness about product M. Company X has 10 websites within its network that have a target audience that would be likely to click on the ad based on data from previous campaigns with companies selling similar products. However, company A is only interested in running on two websites.

Complication

The media conglomerate would like a model saved in a pickle that they can use to test on data from other websites to see which ones are likely to have the highest number of clicks. This would ensure that they run the ads on the top 2 websites where they are likely to drive the most clicks to the site and therefore secure the full amount of $40,000. If the goal is not met by this campaign company X will not secure the full amount and will potentially lose any future business with Company A estimated at a minimum of $300,000 per year.

Executive Summary

- The goal was to build a model with an accuracy above 95%, the model presented in this project has an accuracy of 97.25%
- Hyperparameter tuning was useful in improving model accuracy
- 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

Data Set Characteristics and Information

CHARACTERISTICS

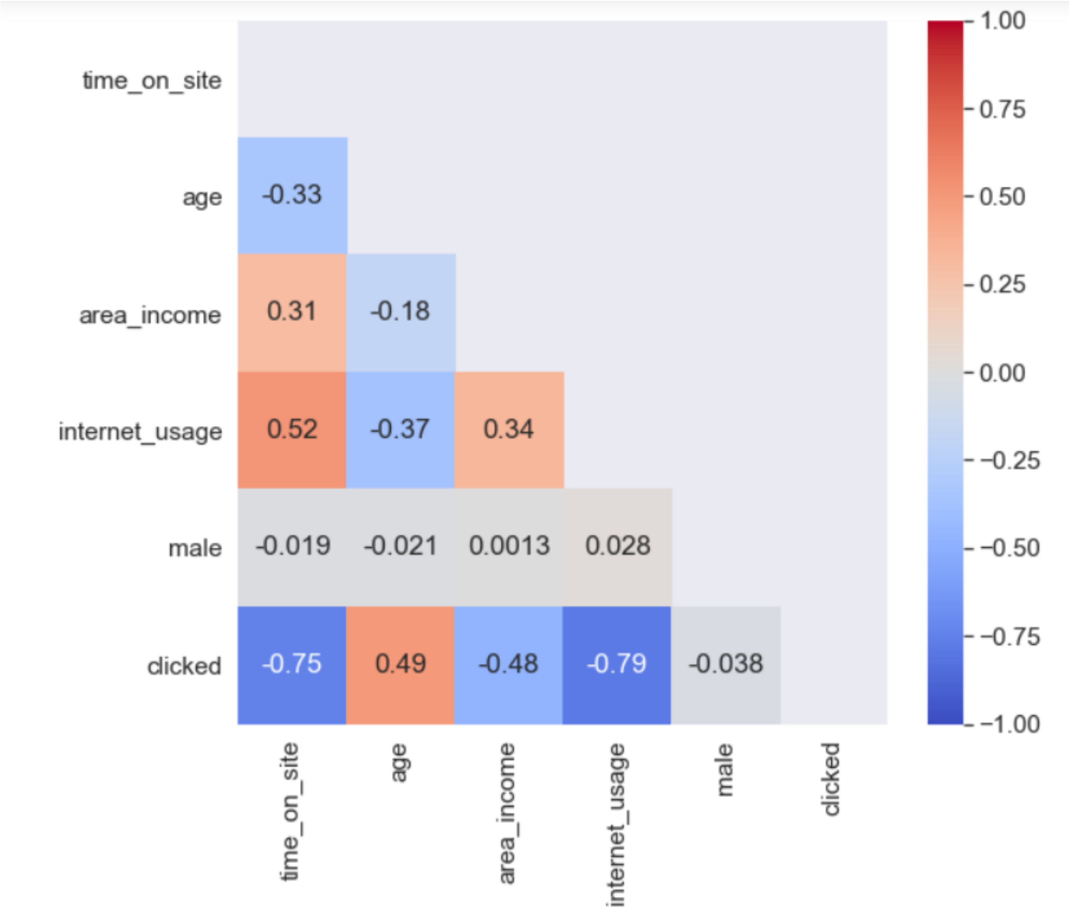
- The dataset used for this project has 1018 rows and 10 columns.
- There were 4 missing values in the target variable column and 7 duplicated values.

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
- **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 (Target Variable)

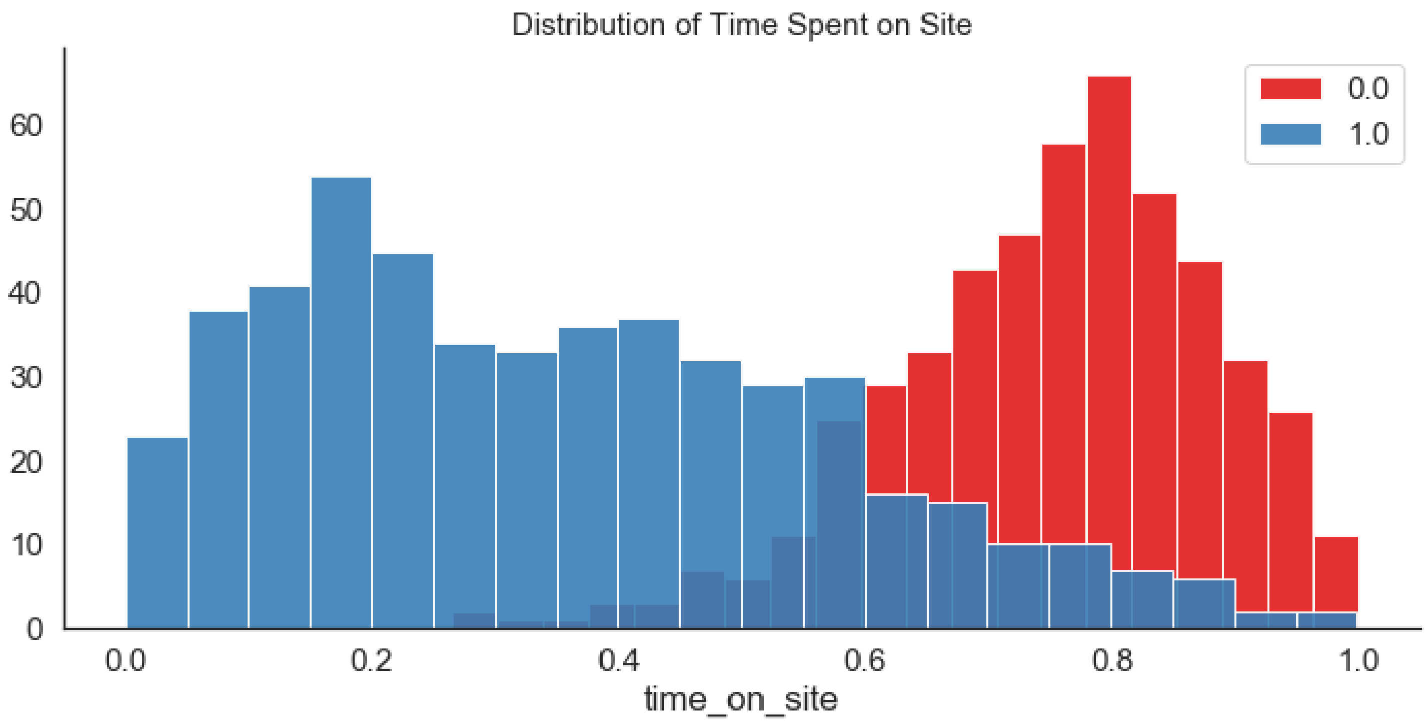
EXPLORATORY DATA ANALYSIS

Internet Usage and Time on Site highly correlated with target variable



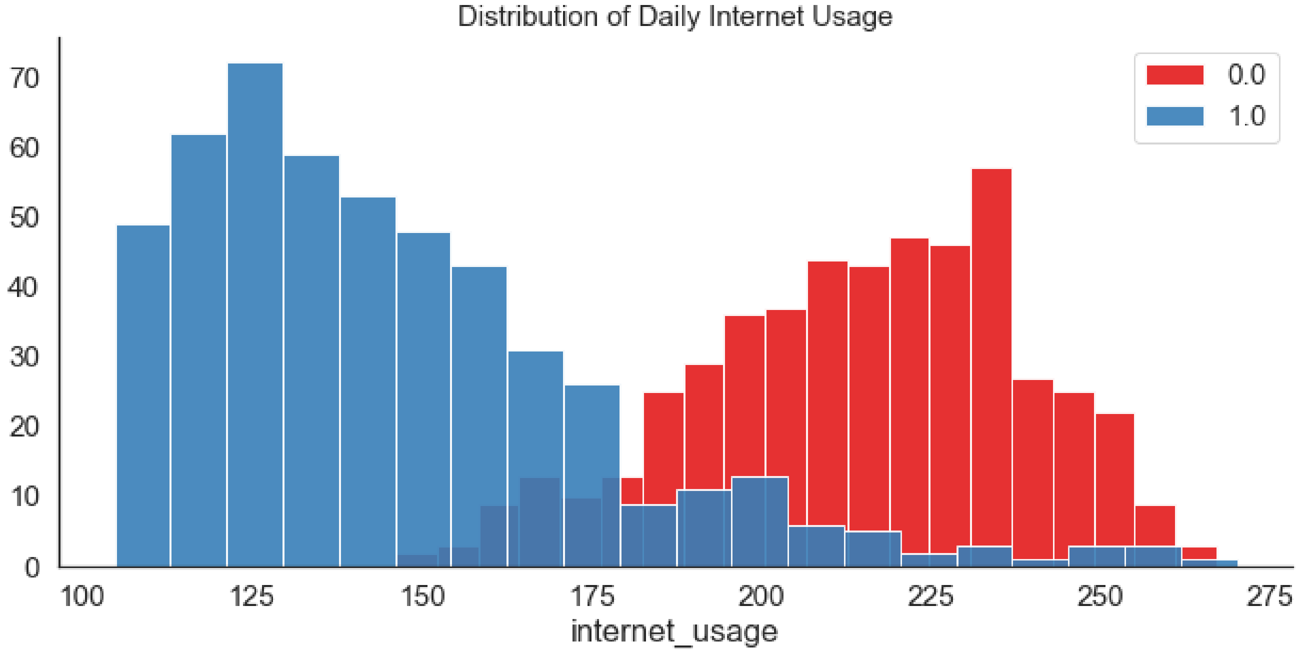
- There is a strong correlation between the target variable and other numerical features except for the gender
- Time on site and internet usage have the highest negative correlation with the target variable
- Age has a positive correlation with the target variable

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



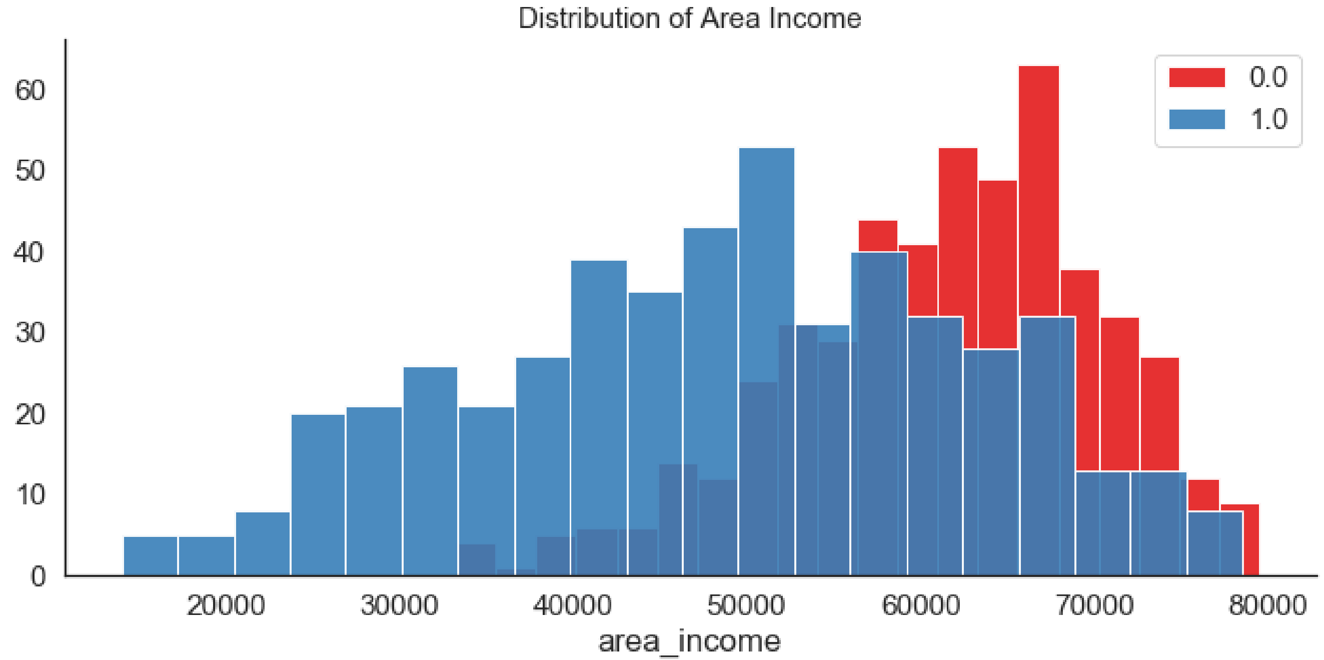
- Time on site is likely to be an important feature in our model since there is a good distinction between the people who clicked on the ad and those that did not
- The avg time that the consumers spent on this site was 65 mins with the minimum time based on this dataset as 32mins

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



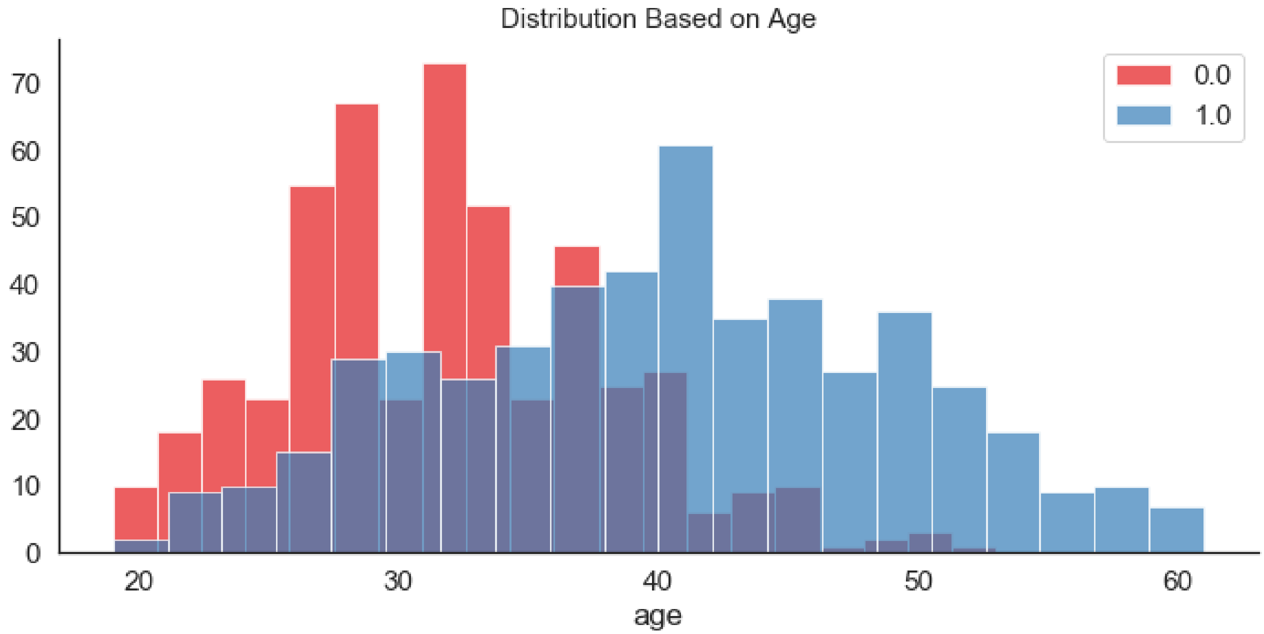
- The avg daily internet usage for the consumers on this dataset is ~3hours. The minimum amount of time spent online based on this dataset was ~24mins
- Consumers who spent less time on the internet were more likely to click on the ad
- There's a good distinction between consumers who clicked on the ad and those that did not and therefore there's a high likelihood that internet usage will be an important feature in our model

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.This was also reflected in the correlation chart
- A likely explanation is that the consumer living in a higher income area does not resonate with the ad

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 B
- 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 this dataset that clicked on the ad is about 50%, 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 those values. 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 this 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 eliminated

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 model was evaluated based on its accuracy and F score
- Initially, accuracy was at 95.50%, the model was tuned and that improved accuracy to 97.25%
- The initial F score was 95.77%, after hyperparameter tuning this increased to 98.25%

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 data could be added based on availability. 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 important factors were age and income. The 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

- Since the different websites have high traffic, it would be very useful to test this model on all their data and select the top 2 websites. Most websites make over 100,000 impressions and therefore with a model with a 97.25% accuracy, we should be able to acquire more than 1600 clicks.

Next Steps & Improvements

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

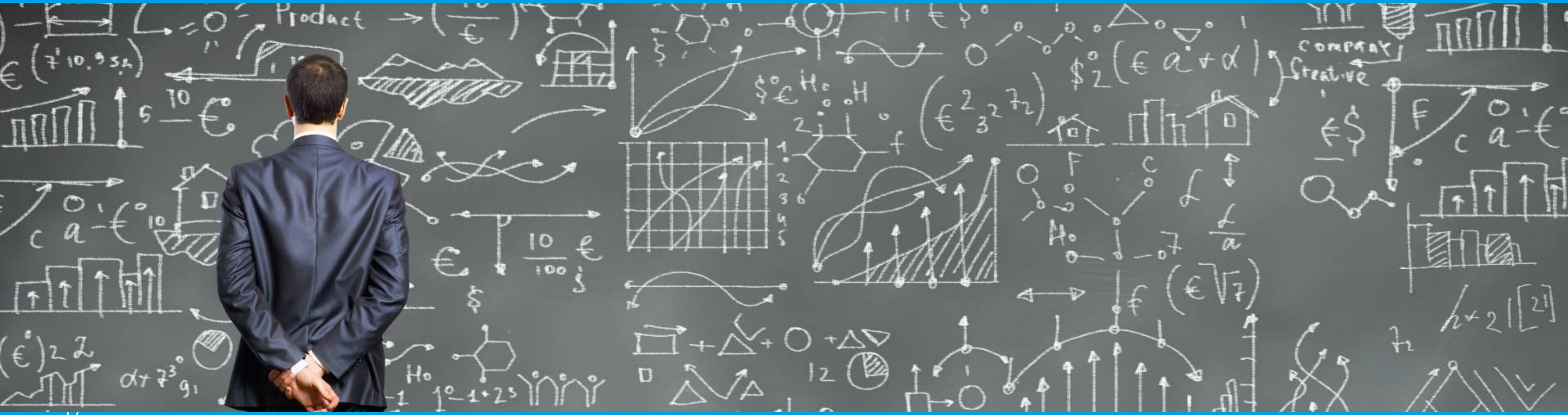
Project/Approach Improvements

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 users resonated towards

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