Advertisement Success Prediction

Employability test

Problem Statement

- The holiday season is just around the corner—Christmas trees have been decorated, lights and wreaths hung, streets all decked up, Santa costumes rented out, and holiday cards in the mailbox.
- Because of holiday cheer, retail brands, big and small, want to earn considerable profits, and therefore, are investing significantly in advertising. These brands have approached an advertising agency to plan and execute ad campaigns that will help them increase the footfall in their stores.
- You have been hired by this advertising company to assess the revenue that can be generated by a proposed ad. Based on the demographic information provided, you need to predict whether the revenue generated will cover costs to produce and air the ad(Whether there will be a net gain from an ad or not)
- This will help guide decision-making for the firm, as they will want to pursue ads that are likely to generate a net gain for their clients— thereby boosting the advertising firm's reputation.

Dataset Characteristics

- The training dataset contains 19536 records and the test dataset contains 6512 records. Following are the features of the dataset

Feature	Feature type	Description
UserID	Categorical, Nominal	Unique id for each row
ratings	Numerical, float	Metric out of 1 which represents how much of the targeted demographic watched the advertisement
airlocation	Categorical, Nomial	Country of origin
airtime	Categorical, Nominal	Time when the advertisement was aired
average_runtime(minutes_per_week)	Numerical, Integer	Minutes per week the advertisement was aired

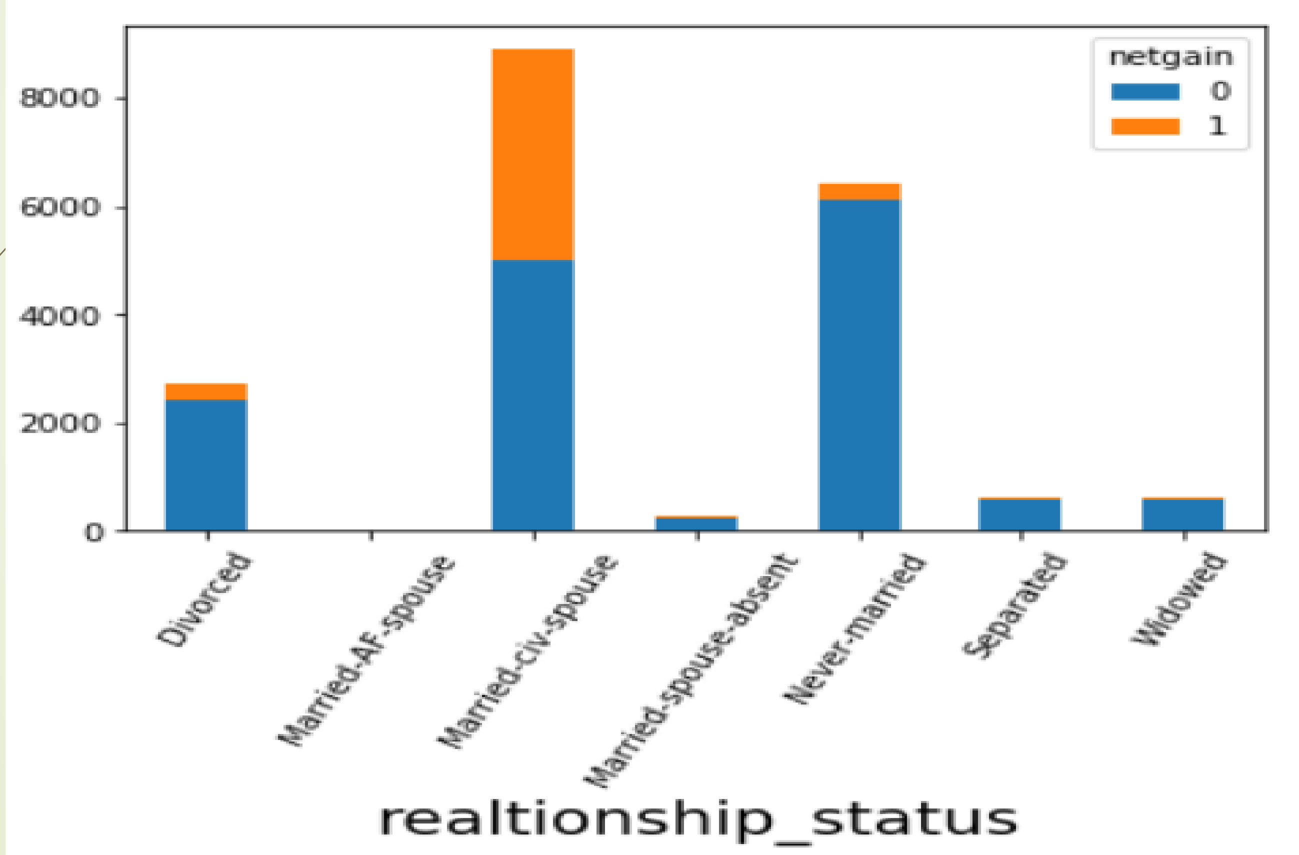
Feature	Feature type	Description
targeted_sex	Categorical, Nominal	Sex that was mainly targeted for the advertisement
genre	Categorical, Nominal	The type of advertisement
industry	Categorical, Nominal	The industry to which the product belonged
relationship_status	Categorical, Nominal	The relationship status of the most responsive customers to the advertisement

Target variable	Target type	description
Netgain	Numerical, Classifier	0 – There is no netgain ; 1 – There is Netgain

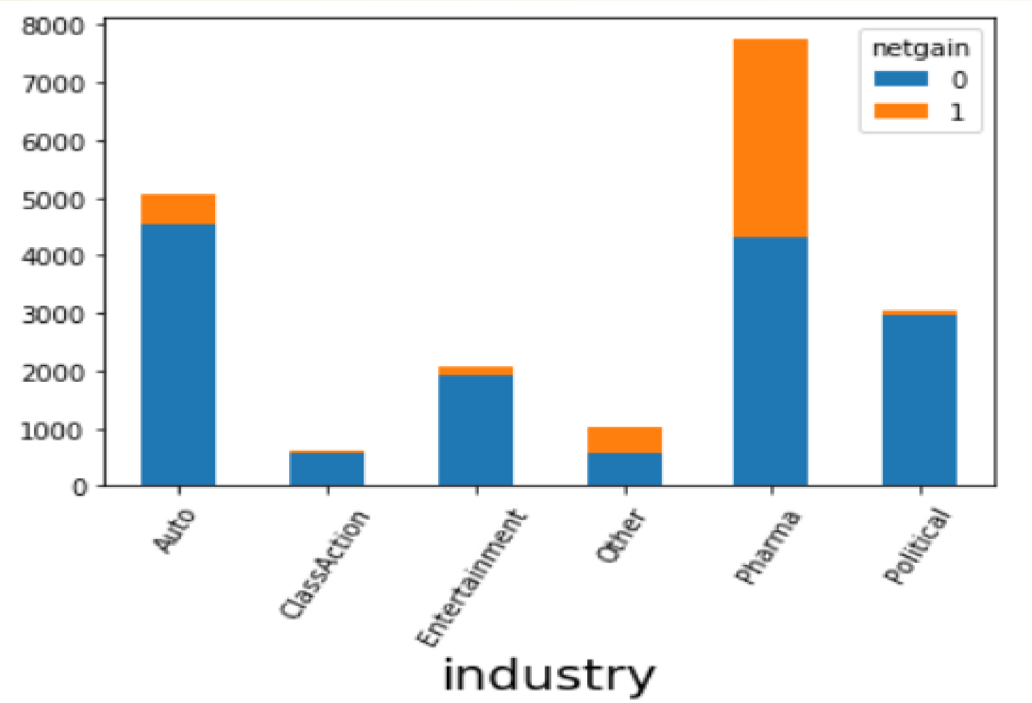
Exploratory Data Analysis

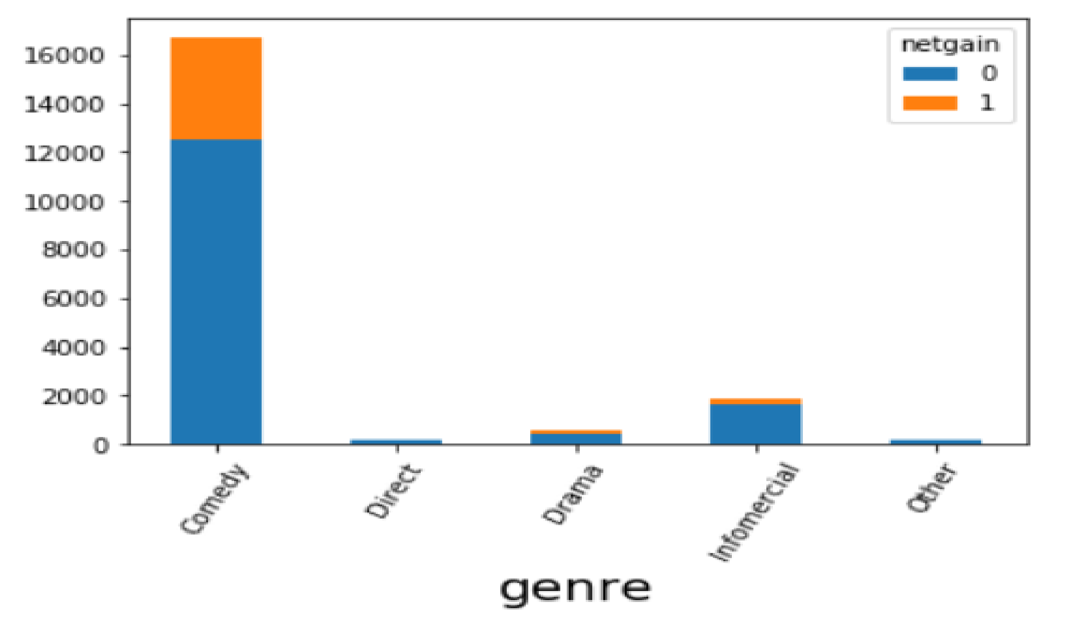
- When checked for missing values, There were no missing values found.

Feature and Target class Analysis of categorical data:



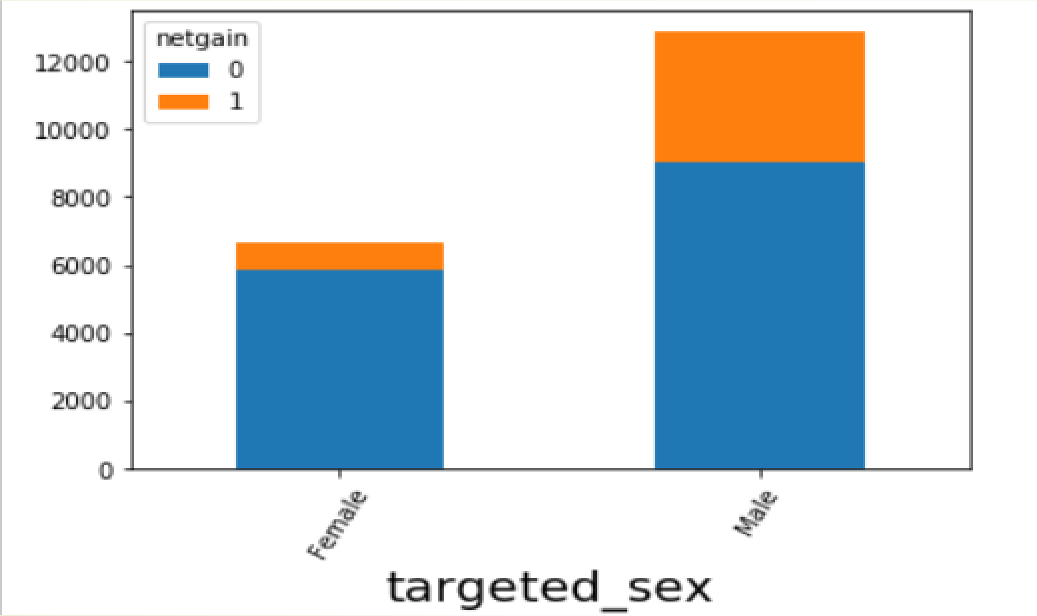
- 50% of the data comprises of the users married to a civilian spouse and the next major relationship status is being never married.



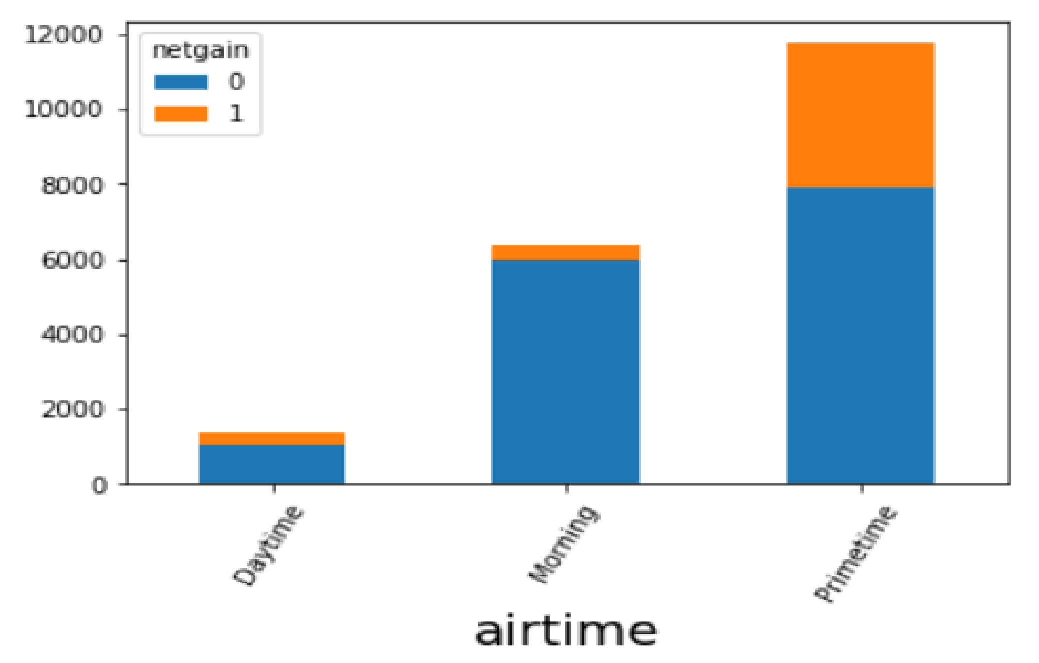


- Almost 40% of the users belong to Pharma industry and automobile, political, entertainment industries being 50% of the data.

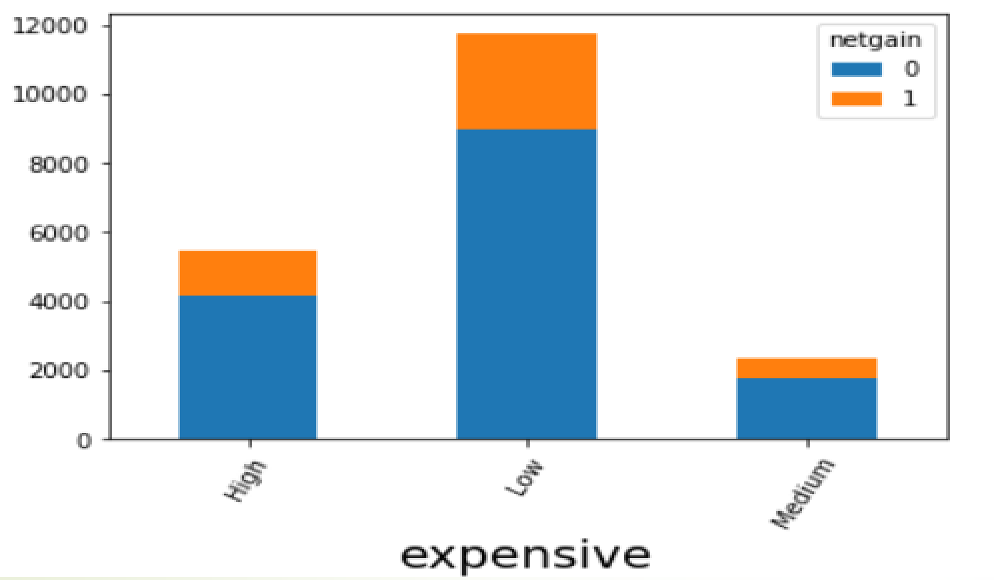
- 90% of the advertisement genres is comedy and the other genres being infomercial, Drama, Direct in the 10% of the data

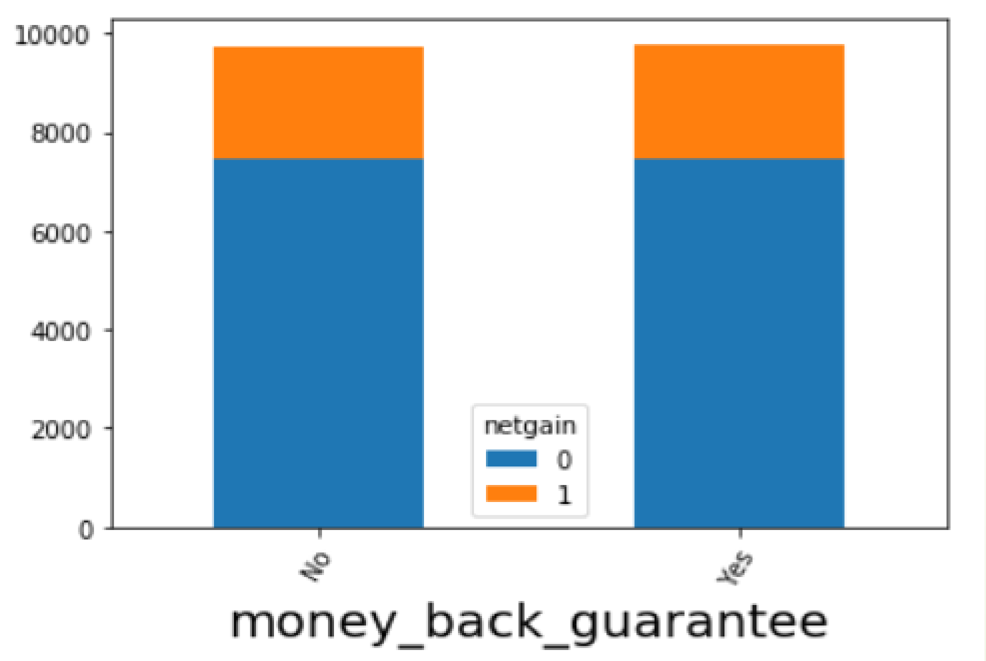


- The maximum targeted sex is male i.e. 70%, Female being only 30%



- 60% of the advertisements are being aired during the prime time, 30% during morning and only 10% during the day time.



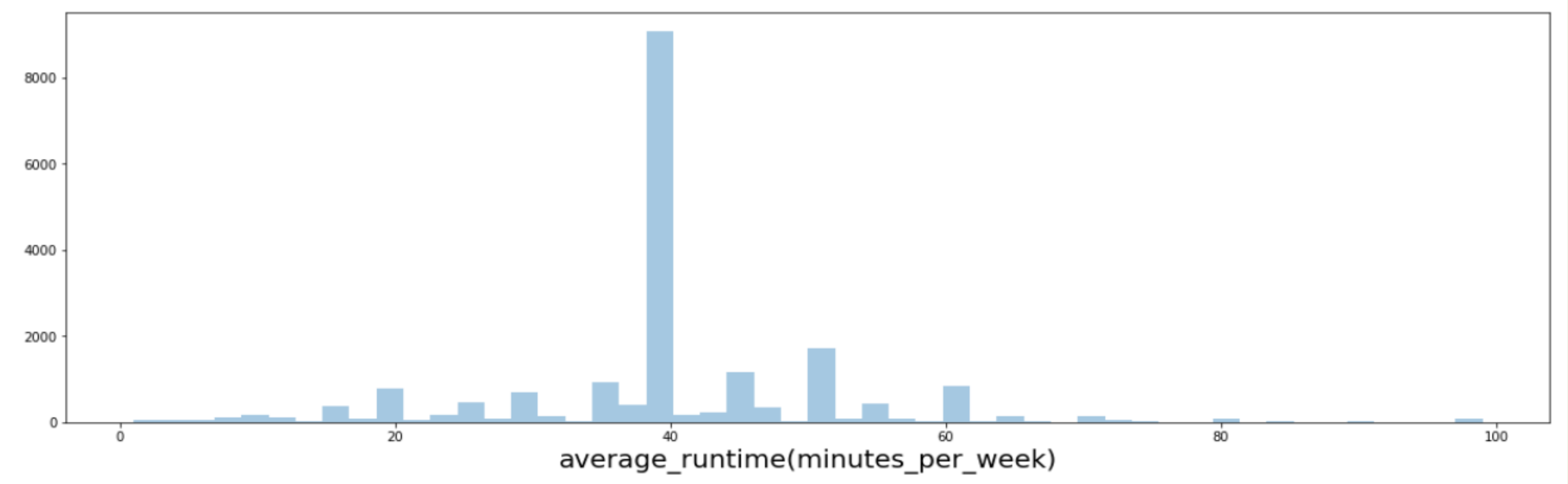


- Almost 60% of the ads are less expensive, 30% being highly expensive and 10% being medium expensive.

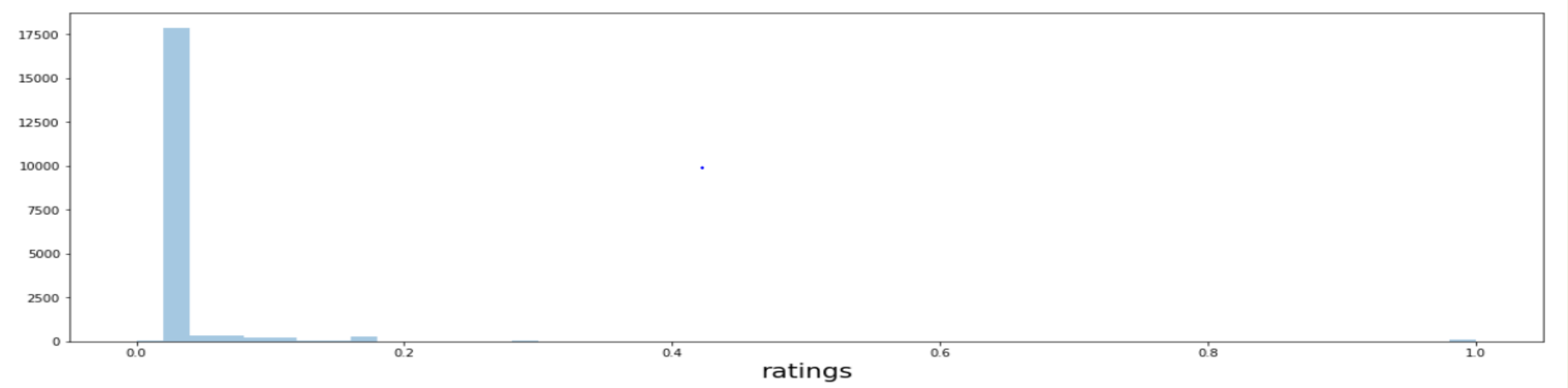
- Half of the ads are having money back guarantee and the other half of them dont.

Univariate Analysis of Numerical data

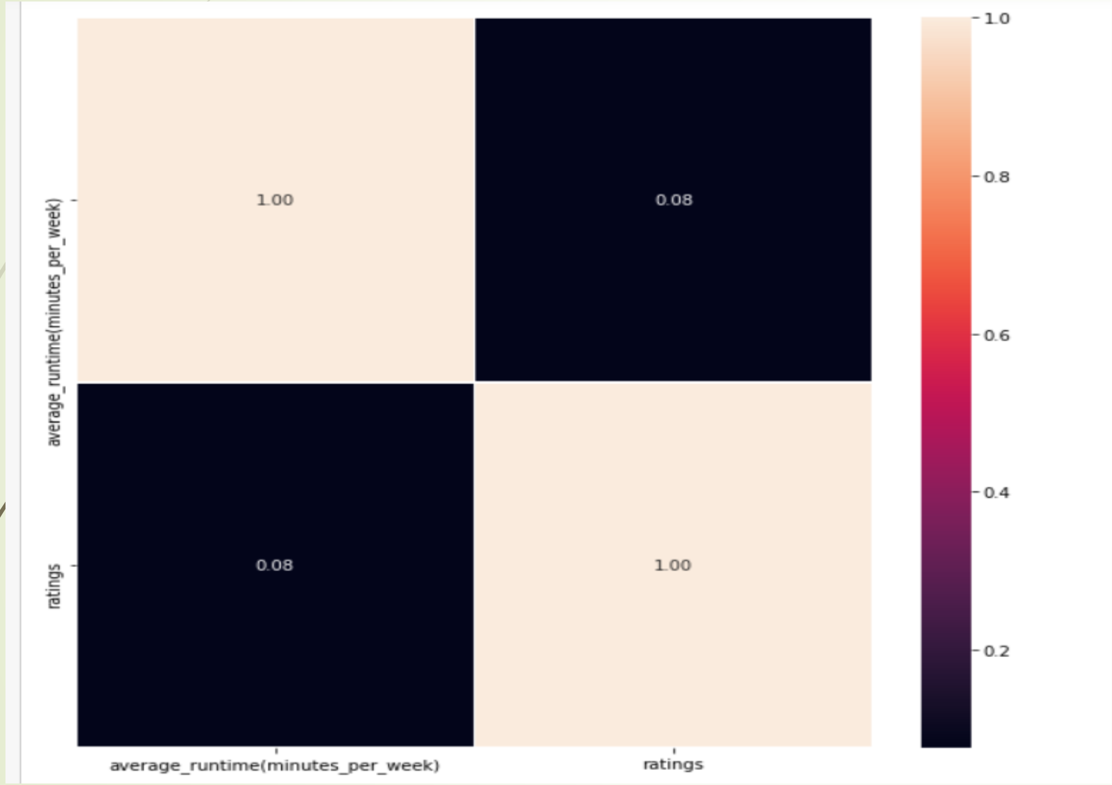
- Maximum number of advertisements are between the runtime 35 – 60 mins per week



- Most of the ratings have the value 0.0274



Feature Correlation



- Didn’t observe much correlation between the below numeric features. Hence decided to retain both the columns.
- Average_runtime(minutes_per_week)
- Ratings

Encoding and Data cleaning

- Label Encoding was done on all the categorical features.
- Since the values in the column *average_runtime(minutes_per_week)* are highly varying. Normalised this column data using StandardScaler
- Since more than 90% of the airlocation data is USA. This column data didn't seem to be useful to be retained. Hence dropping this column.
- Converted the UserID unique identifier column to a numeric column using split function before modelling.

Base line model performance:

Model	F1 Score	Precision	Recall	AUC score
Logistic Regression	0.0967	0.335	0.056	51.6%
Decision Tree Classifier	0.501	0.503	0.498	67.16%
Random Forest Classifier	0.515	0.586	0.46	67.89%
Gradient Boosting Classifier	0.542	0.668	0.456	69.25

- The Best model out of these is Gradient Boosting model.
- Since this is a imbalanced class, Sampling would bring out better performance.

Under Sampling and Over Sampling:

	F1 score	Precision	Recall	AUC score
Tomek Links	0.533	0.55	0.51	69.1%
Random Undersampler	0.62	0.48	0.86	78.66%
SMOTE	0.579	0.458	0.786	74.6%

- Random Sampler is giving the best performace out of all sampling methods.
- Hence selecting Grandient boosting + random sampler for prediction

To improve predictability

- Augument existing data with more data: Like the arilocation column with more countries included would have made this feature valuable.
- Ratings values looks to be some default value selected. If genuine ratings were provided this would increase the performance.
- Adding more features like channel details where ads are posted, Frequency of ads getting updated, Is the channel being changed while the ad is being played.

Thank you