Description

This code predicts whether a comment is an ad or not. I have used machine learning to classify comments. The code is written in python. The technique used here is **TF-IDF** (term frequency-inverse document frequency) for text mining of the comments followed by a simple XGBoost on TF-IDF. TF-IDF is an efficient and simple algorithm for matching words in a query to documents that are relevant to the query. Xgboost is quit popular on Kaggle and needs no introduction. XGBoost is a library for developing fast and high performance gradient boosting tree models. That XGBoost is achieving the best performance on a range of difficult machine learning tasks. Since the dataset mostly contains Text data, I have personally tested the result using **Multi class Log-Loss evaluation metric** by splitting the train dataset into 2 sets. This particular approach **scored 0.279 logloss** which is the best result I got while compared to other approaches like

1. Fitting a simple logistic regression on TFIDF = 0.355 logloss
2. Fitting a simple Naive Bayes on TFIDF = 0.322 logloss
3. SVM = 0.289 logloss
4. Counter vectorizer followed by logistic regression = 0.289 logloss
5. Fitting a simple Naive Bayes on Counts = 1.498 logloss

Also, The **AUCROC score** of my approach was **0.96**