# Impression Throttling

## Background

We currently receive approximately 35 billion mobile ad impressions a day across mobile web and mobile app. We receive a non-zero bid response from DSPs on around 5% of these impressions. We eventually monetize less than 200 million impressions i.e. we currently monetize and serve a creative on around 0.5% of the impressions that we receive. From a bids perspective, these ~35 billion mobile impressions result in over 225 billion bid requests to DSPs. Out of these 225 billion bid requests we receive only around 7 billion non-zero bid response from DSPs and less than 200 million end up winning the auction and being served to the user.

These unmonetized bids and impressions load our infrastructure as well as those of the DSPs and cost us. If we can efficiently and accurately determine the impressions to throttle then we can send the most relevant impressions to the DSPs thereby making efficient use of their infrastructure and also managing their QPS expectations effectively. It’ll also help reduce the load on our infrastructure and also reduce latency as we’d be able to reject impressions quickly at our end without having to wait for a response from the DSP.

Goal

Apply machine learning to determine which impressions should be throttled at our end while ensuring minimal impact to revenue.

## Input

Input data set that can be used for training and validation can be downloaded from [this one drive link](https://pubmaticinc1-my.sharepoint.com/:f:/g/personal/dinesh_kandhari_pubmatic_com/EqgnZhcyn1RCo3-NHimlCMYB0-EqLYnBnOrFfZz_n9nSsA?e=TImzV4). This data set corresponds to a couple of hours of logs for a limited set of publishers. Data for each hour is present in a separate directory and has one zip file for each publisher that can be used for training purposes.

## Data Description

Each row in the TSV file corresponds to an impression and contains the following columns

* Timestamp – Unix timestamp when we received the impression
* Hour of the day when we received the impression
* Publisher Id
* Site Id
* Ad Id
* Ad Size Id
* Ad Type (Video = 13/ Native = 12 / Audio = 14, Banner = others)
* Platform Id (pfi) configured in the system
* WURFL platform id (wpfi) – Platform as determined bu WURFL from the user agent
* Country Code – two-digit country code representing the country from where the impression originated
* Device Id (mob.id) – Hashed device id corresponding to the mobile device
* Platform specific device id (mob.pi) – iOS or Android specific device id
* UDID Type (mob.ut) – IDFA, Android Id, IDFV etc
* Hash Type (mob.uh) - UDID hash type(0-Unknown,1-Raw,2-SHA1,3-MD5).
* Make of Mobile device (mob.mk)
* Model of Mobile device (mob.md)
* OS on the mobile device
* Is javascript enabled on the mobile device (mob.je)
* Location data (mob.lo) – latitude and longitude of mobile device
* Source of location data (los) – IP address, GPS …
* App id (mob.ai) – Application id on the exchange
* App Bundle (mob.ab) – eg. com.foo.mygame
* App Bundle derived (dab) - App bundle extracted from store url
* Ad orientation (oi)
* Device orientation (di)
* Mobile Connection Id (con)
* Bid sent to publisher (pp) – Bid value that PubMatic sent to the publisher
* Campaign id to which the winning bid corresponds to (wcid)
* Verified – Indicates if the bid won in the publisher auction and if PubMatic served the creative to the user. Value of 1 indicates PubMatic won the auction and served the creative to the user
* Passback reason (pb) if the impression was passed back

You can find more details related to these fields at <https://inside.pubmatic.com:8443/confluence/display/CC/AdServer+JSON+Log+Format+Specification>

Please feel free to enrich the data set as per your needs.

Note that it is more important for us to not throttle impressions that will be monetized or minimize this number i.e. protect revenue and it is fine to not throttle some impressions that will not be monetized.

**Test data set** that would be used to measure the accuracy of your trained and final ML model can be downloaded from [this one drive link](https://pubmaticinc1-my.sharepoint.com/:f:/g/personal/dinesh_kandhari_pubmatic_com/Eosni6GJZptLo4D_c8hm3T4BaT4DMMXKJ6tQXrfbI79fzg?e=QEEVOg). This data set corresponds to one hour of logs for the next day for the same set of publishers that were used for training and follows the same format as the data that was used for training the model.

Please do not use the test data set for training nor tuning the model in any way. Please use this only to present the results around how well the trained model did at identifying impressions to throttle and more importantly at identifying impressions that should not be throttled.

## Output / Deliverables

* A **Spark ML** model that predicts if an impression that we receive must be throttled or not. Input can be any of the features of your choice except for fields like pp, wcid and verified from the above set of columns.
* F1 score, precision, recall, fall-out, accuracy and the confusion matrix for the above model corresponding to the test data set. Please share the corresponding ROC curve and AUC as well. Please use data for multiple publishers to validate your model.
* A brief description of the proposed solution & it’s pros and cons. Kindly call out the hardware requirements and time needed for training the model and generating predictions / inferences
* Details of any other models and libraries that you tried that performed better than the tuned Spark ML model

## Part 2- Price Prediction

For the impressions that are monetized (verified = 1), predict the closing price (pp) and winning campaign id (wcid). Here you can use all the features from the training set except for the last three columns i.e. pp, wcid and verified. Please share metrics like RMSE while predicting the closing price (pp)