Estimating gender in a mobile environment

Adello

December 3, 2018

1 Problem description

The programmatic advertisement industry offers a multitude of different audience targeting options, mostly required and valued by branding advertisers. There are traditional direct publishers, which display ads within a specific publisher environment. They depend on publisher centric audience research (e.g. questionnaires) or pure common sense to determine needed audiences for future campaigns. In contrast, there is a technical audience targeting approach, which is enabled by a purely programmatic set-up. It connects a big number of publishers and uses the power of algorithms to predict matching audiences in real-time, independent of the publisher environment. Consequently, targeting requirements within the modern online and mobile display world are more precise than ever. However, suppliers of potential advertisement space (mobile websites or applications) are more diverse and unreliable than traditional approaches like TV or print magazines, which are using market agreed audience data to predict and bill audience reach.
Within mobile, millions of applications and websites offer their users' screen-real estate as advertisement space on an impression-by-impression basis. Specialized solution providers, so called demand side platforms (DSPs) are responsible for executing advertisement campaigns on a technical basis and for delivering on the promise of hyper-precise targeting without alienating end-users. At the same time, strict regulations need to be followed to ensure the privacy and protection of personal data of end-users.
In this whitepaper, we describe an algorithm to identify and target end-users based on their gender in a true big-data environment. We are utilizing a large number of data points originating from hundreds of millions of devices and thousands of applications daily. We use this data to derive a metric related to the probability of each individual device belonging to a female or male person[1]. While such approaches are not new, the implementation in a real-time, big-data environment is challenging and requires continuous investment in data-science know-how and hardware set-up. Consequently, this whitepaper ends up with a claim for a neutral validation solution to justify those investments from AdTech companies and therefore higher CPM prices for the industry.

2 Base Data

Adello is receiving and processing several billion potential impressions per day. A significant fraction of them are produced by mobile applications and therefore often contain pseudonymized device identifiers like the Unique Device Identifier (iOS) and the Google Advertising ID (Android). Our stated goal is to use this data together with potential gender information to derive insights about the relation of application usage and gender. Therefore, it is crucial to have consistent and reliable data.
The device identifier information can be flawed due technical mis-handling or intentional fraud on the publisher-side. After accounting for those factors, we estimate a number of 5.8×10^8 unique devices in our core markets of EMEA and APAC.
Only a small fraction of 5% of those devices have gender information attached. This gender information originates from a variety of different applications, especially social networks and other applications requiring user-registration.
However, not all of this information is reliable. Some publishers fake this information to increase the value of their impressions; others just fail to produce consistent information. To ensure a high quality of the base data we evaluate the reliability of gender information based on the consistency of the information. We both estimate the intra-rater reliability (consistency within an application) and the inter-rater reliablity (consistency between applications).
This method results in a list of millions of devices with their full history of application usage of the last 30 days and a known gender originating from over 2000 different applications.

[1] We acknowledge the existence of various other gender identities, but the available data does not allow to identify them.

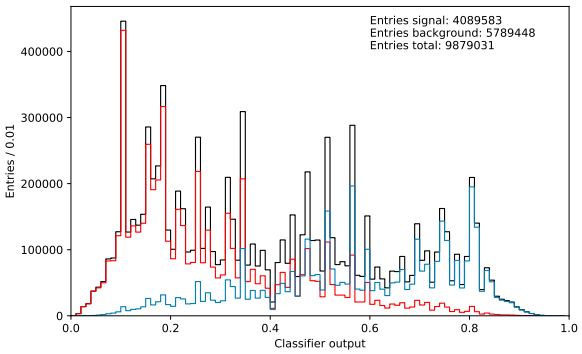

Figure 1: Distribution of classificator output for devices with known gender. The blue, red, and black curve correspond to female, male, and total distribution, correspondingly.

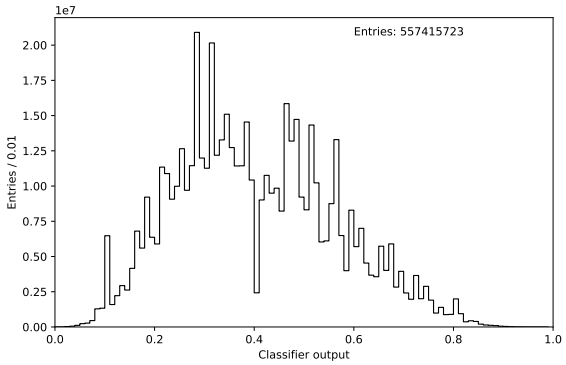

Figure 2: Distribution of classificator output for all devices.

3 Training and Results

The problem at hand is to determine the relation between an input data set (application usage) and a binary output (gender), thus distinguishing between two known classes based on independent input data. This kind of problem is commonly approached with a supervised machine learning algorithm. In the big-data environment, the Spark ML library provides a rich collection of tools and methods to perform various data transformation and machine learning tasks.
We take the full input dataset as described above and vectorize and normalize it. We then train a multivariate, binary classification model based on the Spark ML package. Further, we optimize meta-parameters of the classification model using a grid-search approach. As scoring function we use the area under the receiver operating characteristic curve.
To visualize the performance of the classification algorithm we plot the output of the classificator both for female and male users and study the corresponding precision and recall curves. The plot can be found in Figure 3.

There are two sub-structures observable: A lot of very narrow peaks and a dip in the distribution around 0.4. The peaks are related to the usage of individual applications. Many users are only observed using one application and therefore all of them are assigned the same classificator output. In this way, popular applications materialize as peaks in the output distribution.
In contrast, if users are seen without any meaningful application, the classificator estimates the prior, thus the fraction of female users in the training data sample. Since this output does not contain any additional information with respect to simple guessing, we remove those devices. This discards around 5% of the devices and results in the dip around 0.4.
In the productive use, we want to apply this classificator to all devices, including those with unknown value. But in that case, the overall distribution changes as can be seen in Figure 3. This change is the result of the difference in the datasets. The training dataset only comprises of devices which are observed on applications supplying gender information. The full dataset also contains devices which are not seen on said applications. This inherently results in a different distribution.

Since we want to be able to give a reliable estimation of the precision of a selection, we reweight the

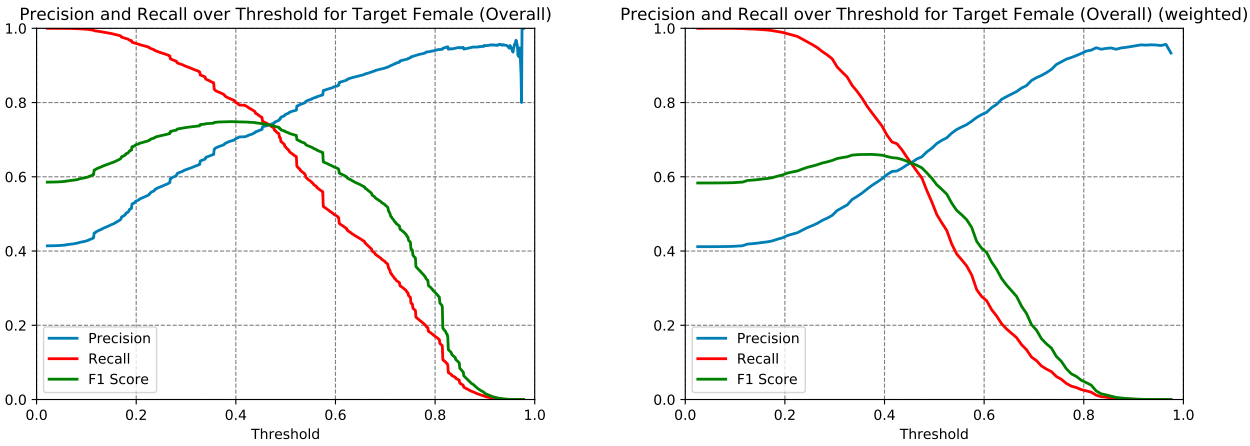

Figure 3: Precision, recall, and F1-score values over the threshold on the classificator output for the unweighted (left) and weighted (right) distribution.

distribution. We take the fraction of female devices for each individual bin of the histogram in Fig. 3 and re-weight the precision and recall values using the full-data distribution in Fig. 3. The difference in precision and recall can be seen in Figure 3. This process is performed separately for each country in our core market.

4 Conclusion

We conclude, that we are able to provide a targeting option based on the gender most probably related to the corresponding device. By utilizing both big-data and machine-learning technology, we can perform this task on a regular basis for hundreds of millions of unique devices per day. However, using existing technology components, while challenging enough, is not sufficient to produce a meaningful result. A thorough understanding of the input data and corresponding sanity and consistency checks are an imperative. In addition, incorporating known differences between training and production data-samples is crucial to produce reliable statements about achieved precision. By doing that, we are able to state that we can fully control the achieved precision up to 0.9 and have a good understanding of the impact on the number of targeted devices.

Call to action:
In our view, above result still needs to achieve wide market acceptance. That highly depends on a neutral and broadly accepted market validation. Currently, there are different efforts by different organizations. Their goal ultimately is the same, to establish such neutral validation systems for the industry, working with established market research companies such as ComScore or GfK. Most approaches follow a hybrid model by combining panel data and newer data-science approaches to determine i.e. age and gender of each displayed campaign (using a measurement tag on campaign level). From a business perspective, such initiatives are overdue. In a media landscape full of uncertainty and unsubstantiated claims a reliable and agreed upon currency is needed to prove targeting quality (precision) and differentiate price tiers. Adello therefore has always and will continuously support all market initiatives to full extent (example in Switzerland: Swiss media data hub SMDH). Advertisers and media agencies should be able to rely on a neutral and transparent guarantee of what they get for their money. However, the arbitrator cannot be Google nor Facebook, we need neutral organizations with clear standards.