

No Free Hunch (<http://blog.kaggle.com/>)



[\(HTTP://BLOG.KAGGLE.COM\)](http://blog.kaggle.com/) > TALKINGDATA MOBILE USER DEMOGRAPHICS COMPETITION, WINNERS'

INTERVIEW: 3RD PLACE, TEAM UTC(+1,-3) | DANIJEL & MATIAS

[\(HTTP://BLOG.KAGGLE.COM/2016/10/21/A-GUIDE-TO-OPEN-DATA-PUBLISHING-ANALYTICS/\)](http://blog.kaggle.com/2016/10/21/a-guide-to-open-data-publishing-analytics/) [➔](#)

[\(HTTP://BLOG.KAGGLE.COM/2016/10/14/GETTING-STARTED-IN-THE-SEIZURE-PREDICTION-COMPETITION-IMPACT-HISTORY-USEFUL-RESOURCES/\)](http://blog.kaggle.com/2016/10/14/getting-started-in-the-seizure-prediction-competition-impact-history-useful-resources/)

TalkingData Mobile User Demographics Competition, Winners' Interview: 3rd Place, Team utc(+1,-3) | Danijel & Matias

[Kaggle Team \(http://blog.kaggle.com/author/kaggleteam/\)](http://blog.kaggle.com/author/kaggleteam/) | 10.19.2016

3

[\(\[\\[mobile-\\]\\(#\\)\]\(http://l</p></div><div data-bbox=\)](http://l)

[user-](#)

[demogr](#)

[compet](#)

[winners](#)

[interview](#)

[3rd-](#)

[place-](#)

[team-](#)

[utc1-](#)

[3-](#)

[danijel-](#)

[matias/](#)




The [TalkingData Mobile User Demographics \(https://www.kaggle.com/c/talkingdata-mobile-user-demographics\)](https://www.kaggle.com/c/talkingdata-mobile-user-demographics) competition ran on Kaggle from July to September 2016. Nearly two-thousand players formed 1689 teams who competed to predict the gender of mobile users based on their app usage, geolocation, and mobile device properties. In this interview, Kagglers [Danijel Kivaranovic \(https://www.kaggle.com/danijelk\)](https://www.kaggle.com/danijelk) and [Matias Thayer \(https://www.kaggle.com/chechir\)](https://www.kaggle.com/chechir), whose team utc(+1,-3) came in third place, describe their winning approach using Keras for "bag of apps" features and XGBoost for count features. They explain how actively sharing their solutions and exchanging ideas in [Kernels \(https://www.kaggle.com/c/talkingdata-mobile-user-demographics/kernels\)](https://www.kaggle.com/c/talkingdata-mobile-user-demographics/kernels) gave them a competitive edge.

The basics

What was your background prior to entering this challenge?


Danijel: I am a Master's student in Statistics at the University of Vienna. Further, I worked as a statistical consultant at the Medical University of Vienna.



Danijel Kivaranovic

Austria




Joined 3 years ago · last seen in the past day



Competitions Master

[Home](#)
[Competitions \(9\)](#)
[Kernels \(1\)](#)
[Discussion \(8\)](#)
[Contact User](#)


Competitions Summary

 <p>Competitions Master</p>	<p>Current Rank</p> <p>136</p> <p>of 50,113</p>	<p>Highest Rank</p> <p>135</p>	<p>Competitions: 8</p> <p>Solo: 6 (75%)</p> <p>Team: 2 (25%)</p>
	 <p>2</p>	 <p>3</p>	

(<https://www.kaggle.com/danijelk/competitions>)

[DANIJEL ON KAGGLE \(HTTPS://WWW.KAGGLE.COM/DANIJELK\)](https://www.kaggle.com/danijelk).


Matias: I've participated in previous Kaggle competitions and most of my relevant experience came from there. Also I work as an analyst now and previously I worked as DBA/developer. I started doing online courses about 2 years ago through EDX, Coursera and MIT professional education. Over there I got familiarized with machine learning, statistics and tools such as R and Python.




Matias Thayer

Santiago, RM, Chile

Joined 2 years ago · last seen in the past day




 [in](#)



Competitions Master

[Home](#)
[Competitions \(19\)](#)
[Kernels \(6\)](#)
[Discussion \(93\)](#)
[Contact User](#)

Competitions Summary

 <p>Competitions Master</p>	<p>Current Rank</p> <p>82</p> <p>of 50,113</p>	<p>Highest Rank</p> <p>81</p>	<p>Competitions: 17</p> <p>Solo: 10 (59%)</p> <p>Team: 7 (41%)</p>
	 <p>4</p>	 <p>3</p>	

(<https://www.kaggle.com/chechir/competitions>)

[MATIAS ON KAGGLE \(HTTPS://WWW.KAGGLE.COM/CHECHIR/COMPETITIONS\)](https://www.kaggle.com/chechir/competitions).

Do you have any prior experience or domain knowledge that helped you succeed in this competition?

Matias: Past competitions in Kaggle, and also familiarity with doing sql-like manipulations on data

Danijel: All prior experience I had comes from the Kaggle competitions I participated in. The datasets in medical research often have less than 100 observations and one is more interested in statistical inference than in black-box predictions. Of course, advanced machine learning tools are not even applicable to these small sample sizes.

How did you get started competing on Kaggle?

Danijel: I heard of Kaggle a few years ago but started my first competition last year ([Rossmann Store Sales \(https://www.kaggle.com/c/rossmann-store-sales\)](https://www.kaggle.com/c/rossmann-store-sales)). Kaggle was a great opportunity to revise what I have learned, improve my programming skills and especially my machine learning knowledge.

Matias: My first competition was in October 2014. It was “15.071x - The Analytics Edge (Spring 2015)”. It was part of an MIT course through EDX. The competition was a lot of fun and I quickly got addicted to Kaggle competitions. I really like learning different viewpoints on hard problems and Kaggle is great for that.

What made you decide to enter this competition?

Matias: I liked the fact that the amount of data was relatively small which means you can do many experiments on a normal laptop. Also, I really like when the data is not “hidden” and you can think and try different hypotheses about it.

Danijel: The data had to fit in RAM. My PC only has 6GB RAM.

Let's get technical

What preprocessing and supervised learning methods did you use?

Danijel: Around 2/3 of devices had no events and the only information available was phone brand, device model and the binary flag if a device has events or not. So, from the beginning, I started to train separate models for devices with/without events. I only used xgboost and keras for modelling.

I used completely different features for my xgboost models and my keras models which was especially beneficial for ensembling afterwards.

Three types of features were used for xgboost:

1. count how often each category appears in the app list (all apps on the device)
2. count how often each app appears in the event list
3. count at which hour and at which weekday the events happened. And also median latitude and longitude of events.

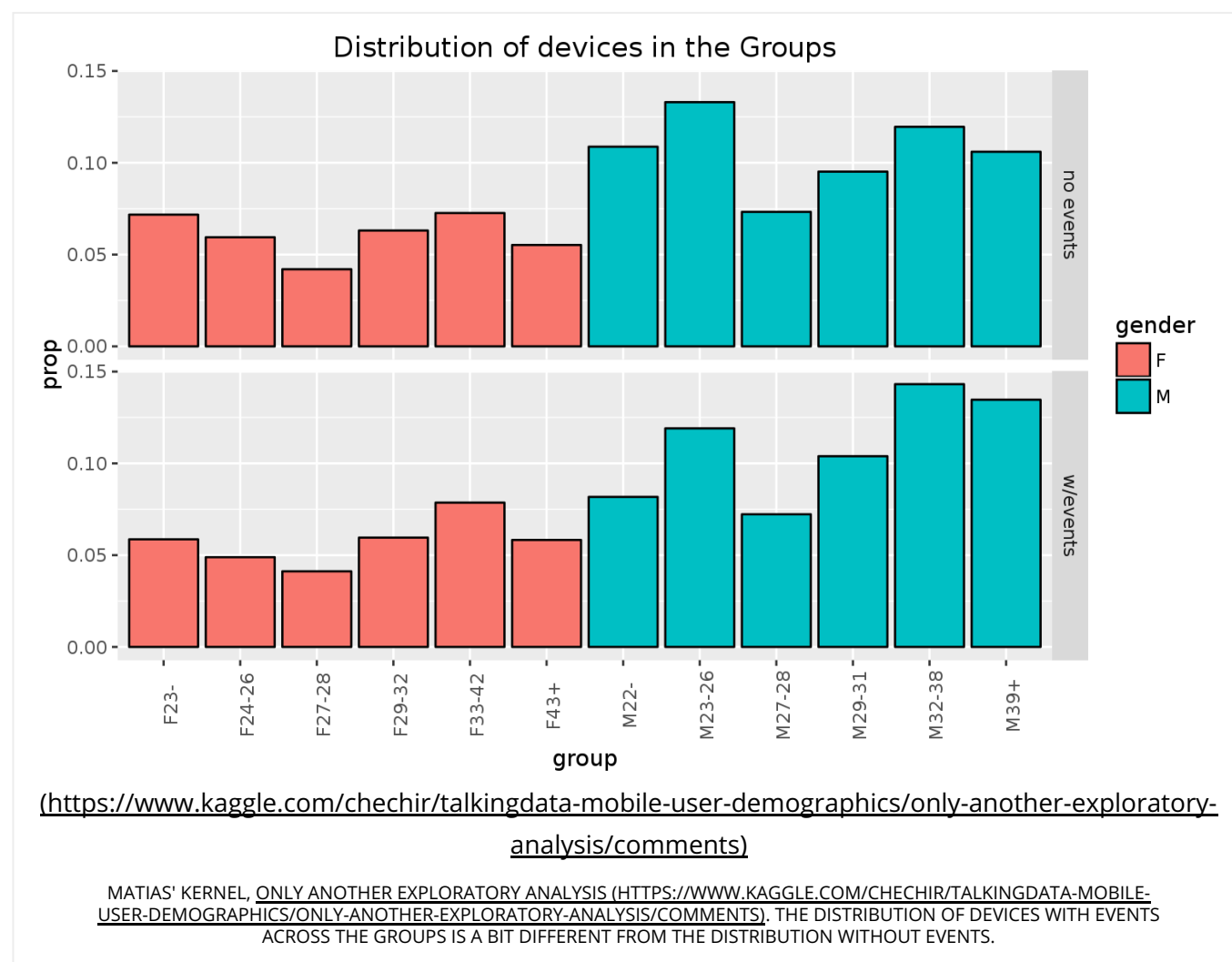
As many other competitors I used the “bag of apps” features for my keras models.

Instead of directly optimizing the logloss, I also tried a 2-stage procedure where I used the gender feature as a meta feature in the second stage. The procedure is:

1. Create the set of features
2. Predict the probability of gender (Stage 1)
3. Use gender as additional feature and predict the probability of age groups (Stage 2)
4. Combine the predictions using the definition of conditional probability: $P(A_i, F) = P(A_i|F)P(F)$ for $i = 1, \dots, 6$ and $P(A_i, M) = P(A_i|M)P(M)$ for $i = 1, \dots, 6$, where the A_i denote the age groups 1 to 6, and F and M denote female and male, respectively.

This 2-Stage procedure significantly outperformed the standard approach for xgboost but was slightly worse for keras.

Matias: I started doing some analysis on the ratios of usage of different apps (in fact that analysis is here: [Only Another Exploratory Analysis \(https://www.kaggle.com/chechir/talkingdata-mobile-user-demographics/only-another-exploratory-analysis\)](https://www.kaggle.com/chechir/talkingdata-mobile-user-demographics/only-another-exploratory-analysis)) and trying with bags of brands, models and labels.



Then I saw the script of Yibo (<https://www.kaggle.com/yibochen>) (XGBoost in R (<https://www.kaggle.com/yibochen/talkingdata-mobile-user-demographics/xgboost-in-r-2-27217>)) and I copied his way to encode everything to 1/0 (even the ratios). Then I started to use a bunch of xgb models as well as a glmnet model, blending them all. I was doing reasonable well (around 20-30 place on the LB) when I saw dune dweller's (<https://www.kaggle.com/dvasyukova>) script. At that time I was trying to learn Keras, so I used her feature engineering and plugged in a keras model. It had a great performance and boosted my score to the 17th position!

I decided to share this Keras script in Kaggle just to get some feedback: [Keras on Labels and Brands](https://www.kaggle.com/chechir/talkingdata-mobile-user-demographics/keras-on-labels-and-brands) (<https://www.kaggle.com/chechir/talkingdata-mobile-user-demographics/keras-on-labels-and-brands>).

```
model=baseline_model()

X_train, X_val, y_train, y_val = train_test_split(Xtrain, dummy_y, test_size=0.002,
random_state=42)

fit= model.fit_generator(generator=batch_generator(X_train, y_train, 32, True),
                        nb_epoch=15,
                        samples_per_epoch=70496,
                        validation_data=(X_val.todense(), y_val), verbose=2
                        )
```

(<https://www.kaggle.com/chechir/talkingdata-mobile-user-demographics/keras-on-labels-and-brands/code>)

MATIAS' KERNEL, [KERAS ON LABELS AND BRANDS \(HTTPS://WWW.KAGGLE.COM/CHECHIR/TALKINGDATA-MOBILE-USER-DEMOGRAPHICS/KERAS-ON-LABELS-AND-BRANDS/CODE\)](https://www.kaggle.com/chechir/talkingdata-mobile-user-demographics/keras-on-labels-and-brands/code) WHICH BORROWS DATA MANIPULATION FROM DUNE_DWELLERS' KERNEL, [A LINEAR MODEL ON APPS AND LABELS \(HTTPS://WWW.KAGGLE.COM/DVASYUKOVA/TALKINGDATA-MOBILE-USER-DEMOGRAPHICS/A-LINEAR-MODEL-ON-APPS-AND-LABELS\)](https://www.kaggle.com/dvasyukova/talkingdata-mobile-user-demographics/a-linear-model-on-apps-and-labels).

And our best model single model for devices with events is just that model with some new features and more layers and regularization. It scored 2.23452 on the LB.

The additional features to this model were:

- TF-IDF of brand and model (for devices without events)
- TF-IDF of brand, model and labels (for devices with events)
- Frequency of brands and model names (that one produced a small but clear improvement)

We merged our teams with Danijel later in this competition, and he was doing something quite different. Together we started retraining some of our models on CV10 and CV5, bagging everything as much as our computers allowed. For the ensemble weights we used the [optim package in R](http://stat.ethz.ch/R-manual/R-devel/library/stats/html/optim.html) (<http://stat.ethz.ch/R-manual/R-devel/library/stats/html/optim.html>) (iterated 3 times) and also built different ensembles for devices with events and devices without events.

When the leak issue was raised we were around 11th to 13th on the LB and we started to look for where it was. After realizing what it was by looking at the files in a simple spreadsheet my teammate Danijel built a clever matching script that, combined with our best submission, allowed us to fight for the top places in those crazy 3 last days. We found also that devices with events weren't taking advantage of the leak, so we only used leak models on non-events devices.

What was your most important insight into the data?

Matias: I found it interesting how the keras model worked out with a high dimension sparse matrix. Also I was really surprised after I opened the train and test sets in a simple spreadsheet. I need to do that more often.

Danijel: As already mentioned, I used two different set of features (count features for xgboost and “bag of apps” features for keras) that performed differently depending on the learning algorithm.

Xgboost: The count features outperformed the “bag of apps” features.

Keras: The “bag of apps” features outperformed the count features. I tried to scale the count features (standard and minmax scaling) but they still could not keep up with the “bag of apps” features which are all one-hot-encoded.

What was the run time for both training and prediction of your winning solution?

Danijel: The best single model takes less than an hour, however, the final ensemble takes a day approximately.

Matias: My best single model (keras) takes about 12 hours in a standard laptop. Mainly because it was bagged 5 times. My other secondary models took between 2-8 hours to run (usually overnight).

Words of wisdom

What have you taken away from this competition?

Matias: At first I thought that sharing my scripts as kernels would make me weaker in terms of ranking, because everyone could see my ideas, but to the contrary my final rank wasn't bad and it helped me a lot to validate things and get really valuable feedback from other users. Also, I added NNET with Keras to my coding library.

Danijel: I learned three things:

1. It is a huge step from Top 10% results to Top 10.
2. I need better hardware.
3. How to install Keras on a Windows PC.

Teamwork

How did your team form?

Danijel: I contacted a few top competitors. Matias was the first who agreed to merge.

Matias: I received an invitation from Danijel and I accepted it.

[TALKINGDATA MOBILE USER DEMOGRAPHICS
\(HTTP://BLOG.KAGGLE.COM/TAG/TALKINGDATA-MOBILE-USER-
DEMOGRAPHICS/\)](http://blog.kaggle.com/tag/talkingdata-mobile-user-demographics/)

[XGBOOST \(HTTP://BLOG.KAGGLE.COM/TAG/XGBOOST/\)](http://blog.kaggle.com/tag/xgboost/)

3 Comments

No Free Hunch

 Login ▾

 Recommend 3

 Share

Sort by Best ▾



Join the discussion...



Carlos R. Cerrato E. • 9 days ago

After finishing my current ML courses on Coursera I'm definitely going to Kaggle competitions. Practice is the best way to learning I think (as oppose to just read or watch videos)

1 ^ | v • Reply • Share ▸



Nipun Agarwal • 10 days ago

Nice to hear this, thanks

1 ^ | v • Reply • Share ▸



Hang Yu • 11 days ago

Great talk!!! Thanks for sharing!!!

1 ^ | v • Reply • Share ▸

 Subscribe  Add Disqus to your site Add Disqus Add  Privacy

DISQUS



<https://www.facebook.com/kaggle>



<https://twitter.com/kaggle>

