**III. Model documentation and write-up**

1. **Who are you (mini-bio) and what do you do professionally?**

Last seven years I am working in the field of mathematical analysis of biomedical signals. The most interesting topics for me are a study of work of human brain and artificial intelligence. I believe that machine learning and artificial intelligence are one of the key technologies of our future. Last 3 years I am working on a Ph.D. thesis which is devoted to the prediction of epileptic seizures based on EEG analysis. Currently, I am working as a research engineer in R&D lab in Ciklum.

1. **High level summary of your approach: what did you do and why?**

My approach is based on [Benchmark: Simple Linear Models](http://blog.drivendata.org/2017/04/28/penguins-benchmark/) – this is a great starting point for this problem. The first what I did is replaced of Linear Regression with XGBoost and minor parameters tuning. After that, I have looked around the preprocessing part and decided to improve function for replacing NaNs in the dataset. Actually, that didn’t give any improvements but allowed to combine different models into an ensemble. Submission generated by ensembling of previous submissions resulted in 5.2132 on the public leaderboard and 4.9605 on the private.

1. **What are some other things you tried that didn’t necessarily make it into the final workflow (quick overview)?**

I’ve tried LightGBM, but the first submission gave much worse result comparing to XGBoost and I had no time and submissions to work around the improvement of this model. Another model which gave low score was one, where I implemented linear interpolation with extrapolation for missing values. These two models were not used in the final ensemble.

1. **Did you use any tools for data preparation or exploratory data analysis that aren’t listed in your code submission?**

No, I didn’t.

1. **How did you evaluate performance of the model other than the provided metric, if at all?**

I had no time to develop additional tools for performance evaluation and some more sophisticated data analysis.

1. **Were there other fields or features you felt would have been very helpful to have? That is, what might the organization want to collect in the future that was not in the given data?**  
   It would be interesting to analyze such data along with weather data – maybe there are some dependencies, which may be used to improve an accuracy of the model.
2. **Anything we should watch out for or be aware of in using your model (e.g. code quirks, memory requirements, numerical stability issues, etc.)?**

-

1. **Do you have any useful charts, graphs, or visualizations from the process?**

No.

1. **If you were to continue working on this problem for the next year, what methods or techniques might you try in order to build on your work so far?**

* Extract and use more features
* Improve handling of missing data
* Generate more diverse models for ensembling

1. **Are you willing to be interviewed for a blog post?**

Yes, it would be great.