Slide 1:

Hello everyone, this is team 18 and that’s our presentation of our Introduction to Data Science project. The dataset for our project was called ‘Zindi User Behaviour Birthday Challenge’.

Slide 2:

The main problem for the Zindi platform as for the website for data science competitions is the leak of users. So, the main goal of the task is to predict whether a person will be an active user of that platform in the future months, understand the reason of leaving and eliminate it. To achieve the first point, it is a nice decision to use data science techniques, that should lead to the quality of the users’ service improvement.

Slide 3:

We started our work with the loading the data. The particular dataframes you can see on the screen. Due to the initial data and targets, there’s the task of the binary classification, so we had to preprocess our data exactly for algorithms, which deal nicely with such kind of tasks.

Slide 4, 5:

After that we began preparing the dataset: one-hot encoding for Logistic Regression and Random Forest models performance. Moreover, we generated several features like ‘account age’ and ‘the sum of activities in the previous month’, because we decided that such kind of things matter in cases when it comes to the activity in the future.

Slide 6, 7:

During our exploration we got some infographics. For example, the heatmap of correlation of the train features and the target, histogram of distribution of observations among the groups and feature importance infographics.

Slide 8:

We decided to check out 3 models: Logistic Regression, Random Forest and CatBoost. According to the received roc-auc and F1 scores we understood that CatBoost gives out the best performance.

Slide 9:

Later we decided to maximize CatBoost’s output, so with the help of Grid Search got the best parameters according to our task. Hence, we got the metrics improvement.

Slide 10:

1. An interesting finding: the target variables components were not fully correct from the start, so we had to reapply the correct values and update targets themselves just in case
2. We used the oversampling for cross-validation and the undersampling for grid-search to understand the model better.

In conclusion: after the data preprocessing, generating features and checking several models, we realized that CatBoost is the best model for that task, and we got quite nice ROC-AUC and F1 metrics.

Slide 11:

Finally, this is our team!

Slide 12:

Thank you for your attention!