Sberbank Data Science Contest
Метрики качества

- **Задача A**. Качество распознавания пола на тестовой выборке измеряется с помощью метрики AUC-ROC[1]. Обозначим число тестовых клиентов за n, пол i-го тестового клиента за $y_i \in \{0, 1\}$, а предсказание для него за $\tilde{y}_i \in [0, 1]$. Тогда качество предсказания вычисляется следующим образом:

$$\text{AUC-ROC} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I[y_i < y_j] I[\tilde{y}_i < \tilde{y}_j]}{\left(\sum_{i=1}^{n} I[y_i = 0]\right) \left(\sum_{i=1}^{n} I[y_i = 1]\right)} \in [0, 1].$$

  Чем выше значение метрики, тем лучше предсказание.

- **Задача B.** Качество прогнозирования объема трат в каждой категории во все дни следующего месяца измеряется с помощью метрики RMSLE[2] со смещением 500. Обозначим число тестовых объектов (все пары категория-день) за n, истинный объем трат всех клиентов в i-й паре за $y_i \geq 0$, а прогноз в ней за $\tilde{y}_i \in \mathbb{R}$. Тогда качество прогноза вычисляется следующим образом:

$$\text{RMSLE}_{500} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\log(y_i + 500) - \log(|\tilde{y}_i| + 500)\right)^2}.$$

  Чем ниже значение метрики, тем лучше прогноз.

- **Задача C.** Качество прогнозирования объема трат каждого клиента в каждой категории в следующей месяц измеряется с помощью метрики RMSLE со смещением 1. Обозначим число тестовых объектов (все пары категория-клиент) за n, истинный объем трат в i-й паре за $y_i \geq 0$, а прогноз за $\tilde{y}_i \in \mathbb{R}$. Тогда качество прогноза вычисляется следующим образом:

$$\text{RMSLE}_1 = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\log(y_i + 1) - \log(|\tilde{y}_i| + 1)\right)^2}.$$

  Чем ниже значение метрики, тем лучше прогноз.

[1] https://en.wikipedia.org/wiki/Receiver_operating_characteristic
[2] https://www.kaggle.com/wiki/RootMeanSquaredLogarithmicError