**Отчет**

**Автор: Ямолдин Александр Алексеевич**

**Задание 1. Выборка по значениям вероятности модели.**

|  |  |  |
| --- | --- | --- |
| Вероятность | Всего в группе | Кол-во таргет 1 |
| [0, 0.1] | 3838 | 44 |
| (0.1, 0.2] | 12245 | 398 |
| (0.2, 0.2] | 10438 | 662 |
| (0.3, 0.4] | 9406 | 926 |
| (0.4, 0.5] | 8584 | 1142 |
| (0.5, 0.6] | 7956 | 1485 |
| (0.5, 0.7] | 7483 | 2116 |
| (0.7, 0.8] | 6166 | 2326 |
| (0.8, 0.9] | 4465 | 2478 |
| (0.9, 1] | 15 | 11 |

**Задание 2. Метрики модели на data.csv**

**rMSE = 0.5433**

**F-мера = 0.447**

**ROC\_AUC = 0.713**

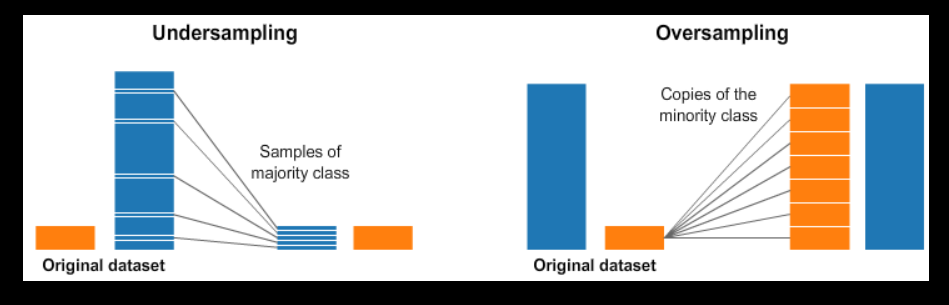
**LogisticLoss = 10.196**

**Ответы на вопросы:**

* **Как можно заполнить пропущенные значения в данных?**
* Борьбу с пропусками можно осуществлять различными способами: убирать строки, в которых пропущены данные (если много данных и мало данных с пропусками), как было в данной задаче.
* Заменять пропущенные значения средним по столбцу
* Лучшим способом является применение KNN Inputera. Суть заключается в том, что мы ищем K- ближайших соседей (похожих строк из датасета) и заполняем пропуск средним по соседям. Минус: очень ресурсозатратный способ

**Что делать, если таргет несбалансированный?**

* Использовать весовые коэффициенты для классов. Class\_weights (этот способ я применял в данной задаче)
* UnderSampling – объекты из мажоритарного выбрасывать и уравнивать выборку до миноритарного класса.
* OverSampling – семплировать объекты миноритарного класса до тех пор, пока миноритарный класс не сравняется с мажоритарным.



**Все методы я применял при решении данной задачи, лучшие метрики получились при использовании весовых коэффициентов для классов.**

**Как бороться с переобучением?**

* Для борьбы с переобучением можно использовать регуляризацию. L1 (lasso) – регуляризация производит отбор признаков, однако не сильно штрафует за выбросы. L2 – регуляризация сильно наказывает за выбросы, но отбор признаков не производит
* Можно тюнить параметры модели и делать модель «слабее», например в бустинге можно ограничивать количество деревьев.

**Найдите потенциальные причины аномалий в наборе данных.**

Основные причины аномалий в наборе данных это то, что данные могут собираться из некачественных источников (напр. Заноситься в БД вручную – человеческий фактор), данные в одного поля могут быть неправильно масштабированы (напр. Вес можно заносить в г, кг, т и …)

В нашем датасете было гигантское количество аномальных данных. Для удаления аномальных данных я использовал метод интерквантильного размаха, в результате датасет был уменьшен в **3.2 раза!** (c 213 927 до 66 786 строк)

**Что делать, если данных не хватает или они плохого качества? Что можно сделать чтобы повысить качество?**

Если данных не хватает, можно попытаться из найти в сторонних источниках, напр. Спарсить с открытых баз данных. Также можно синтетически гинерить данные из имеющегося датасета.

Если данные плохого качества, то их нужно чистить. В данной работе я удалял высоко-коррелированные фичи, убирал из датасета аномалии и нормализовал на единичное мат.ожидание и дисперсию, чтобы централизовать данные.