Курсовой проект мегафон

Для финальной модели проекта используется XGBClassifaer из библиотеки scikit-learn.

Так же была испытана логистическая регрессия из той же самой библиотеки. Результаты моделей были сопоставимы по метрикам, я отдал предпочтение xgboost-у.

Модель была испытана с заменой выбросов по 1.5 интерквартильным размахам, по 90% персентилю, по 95% персентилю, по 99% персентилю. Заменялись на моду, медиану, среднее.

В конечной модели замена не включена совсем.

Так же было испытано применение понижение размерности PCA до 99% онформации данных, что сокращало кол-во колонок датафрейма на 60-90 колонок. Метрики данных испытаний так же были недалеко друг от друга, но при последнем испытании метрика без PCA была на 0.95% выше — поэтому оставил ее для конечного результата.

Так же было реализован вывод в порядке убывания вероятности подключения услуги для всех услуг из датасета. Для файла test.csv это можно увидеть в блокноте predict test\_6.ipynb. Предсказание услуг реализовано путем подставления id услуги в датасет.

Для исключения проблем с оперативной памятью была создана БД MYSQL на облачном сервере линукс и туда заброшены все данные из файла features.csv. Там же реализована страница на сайте, где вводится id пользователя(который присутствует в файле features.csv), дата, услуга. В том числе там выдается ответ на запрошенную услугу, наиболее рекомендуемая и список в порядке убывания всех услуг. Просмотреть можно на сайте http://bainel.ru/predict\_megafon.html.

Конечный ноутбук predict\_ansver.ipynb принимает данные из файла test.csv, выбирает все строки из бд по колонке „id“ и дальше объединяет их путем pd.merge\_asof, и отдается на модель.

В ходе испытаний создавался файл в формате json с номерами удаляемых колонок, порогами разделения и значениями услуг, что бы в дальнейшем их не составило труда найти.