

# Задача прогнозирования оттока пользователей или **churn prediction**

# Цели и задачи проекта

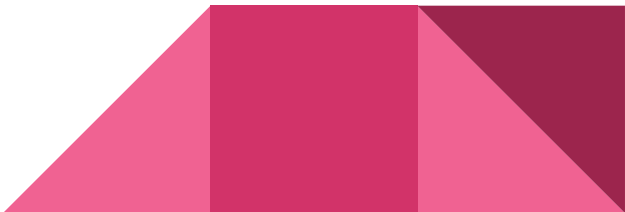
Крупная компания телеком Orange имеет набор данных о пользователях.

**Целью проекта** является построение моделей, с помощью которых возможно заблаговременное нахождение пользователей, склонных к оттоку, то есть к отказу от некоторого продукта компании. С помощью построенной модели можно бороться с оттоком, проводить эксперименты по выявлению причин оттока, и проводить различные акции, связанные с удержанием пользователей.



# Цели и задачи проекта


**Задачи проекта** состоят в следующем:

1. проанализировать и обработать данные, предоставленные компанией;
  2. с помощью проанализированных данных построить и оценить модель, которая позволит в дальнейшем определять заблаговременно пользователей склонных к оттоку;
  3. построить экономические модели для оценки эффективности модели для удержания пользователей.
- 

# Оценка модели и критерий успешности

В качестве основных критериев оценки качества построенной модели можно указать величину ROC\_AUC метрики для построенной модели, а также ее полноту в определении пользователей, склонных к оттоку. Значения этих оценок будет вычисляться с помощью кросс-валидации, кроме того решение будет тестироваться на отложенной выборке.

В качестве критерия успеха модели можно указать успешность кампании по удержанию пользователей, а также размер дополнительной прибыли, который может быть получен с использованием данной экономической модели.




# Техническое описание модели

# Обработка данных

Некоторые признаки в данных являются константными, поэтому это их необходимо удалить.


Кроме этого, обрабатываем числовые признаки. Поскольку в данных имеются пропуски, их необходимо заполнить, например, средним значением.

Нужно дополнительно обработать категориальные признаки. Будем использовать следующую стратегию: будем удалять значения в каждом признаке, если количество объектов, на которых данное значение встречается, меньше 400. Кроме того, применим стратегию OneHot для создания новых признаков: каждому уникальному значению в каждом признаке сопоставим бинарный признак.



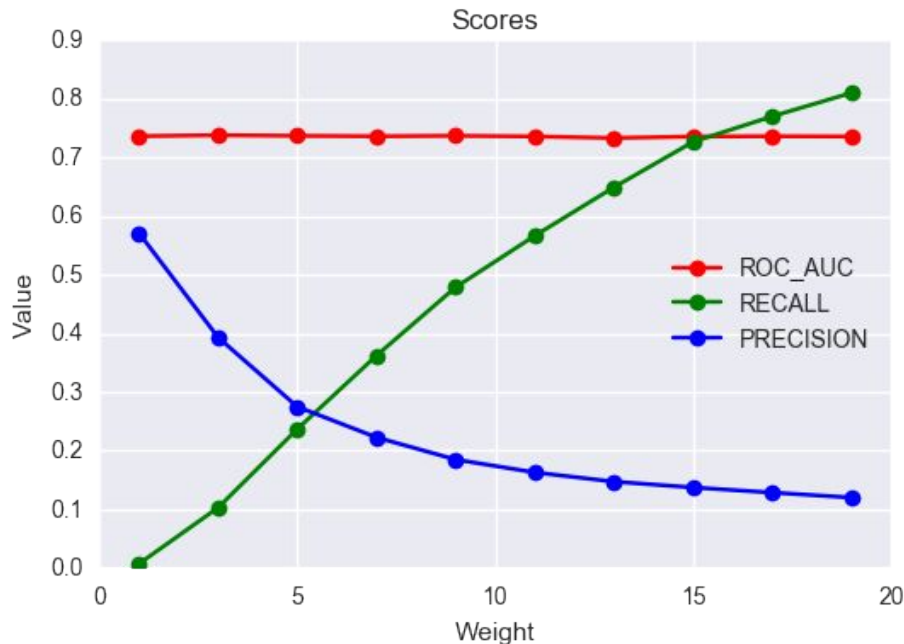
# Построение модели

В качестве базового алгоритма будем использовать `XGBClassifier` (алгоритм градиентного бустинга из библиотеки `xgboost`). С помощью `GridSearchCV` (из `scikit-learn`) были подобраны основные параметры модели, а именно количество деревьев решений, максимальную глубину этих деревьев. Оптимальными параметрами были подобраны следующие значения: максимальная глубина равная 3, количество деревьев равное 100. При выборе данных параметров оценка результатов выполнялась с помощью кросс-валидации с количеством стратифицированных фолдов равных 3.



# Повышение полноты модели

Основной из проблем выборки является ее несбалансированность, что приводит к тому, что обученная модель неудачно предсказывает объекты, склонных к оттоку. Для решения этой проблемы необходимо повысить веса элементов склонных к оттоку. На графике можно пронаблюдать зависимость оценок модели в зависимости от выбранного веса. Таким образом, с помощью весов можно дополнительно улучшить модель и повысить ее полноту и эффективность, при почти неизменном значении оценки ROC\_AUC

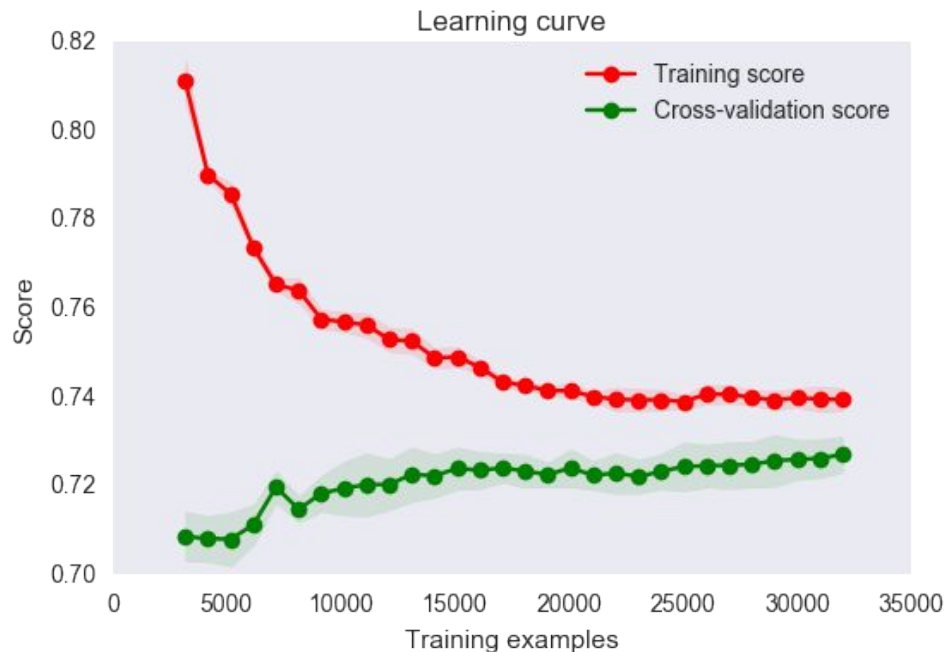




# Оценка полученных результатов

# Оценка модели

Можем наблюдать, что ROC AUC оценка модели, полученная с помощью кросс-валидации, растет при увеличении размера обучения, после чего выходит на плато, и перестает значительно расти, что может говорить о том, что модель отлично обучена

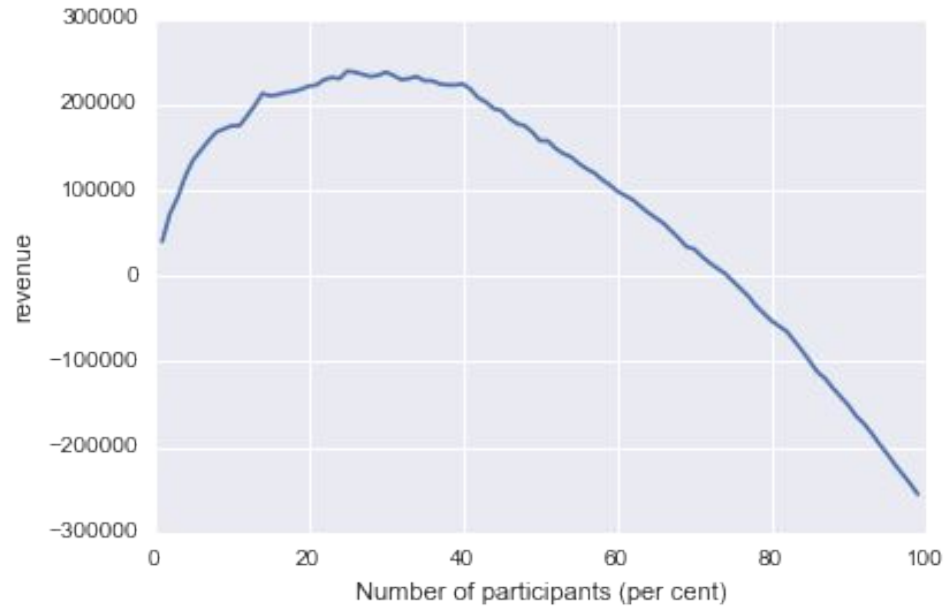


# Экономическая модель

Кроме того, была построена простая экономическая модель по удержанию пользователей, для того чтобы оценить экономический эффект, который может быть достигнут с помощью модели при оптимальных параметрах.

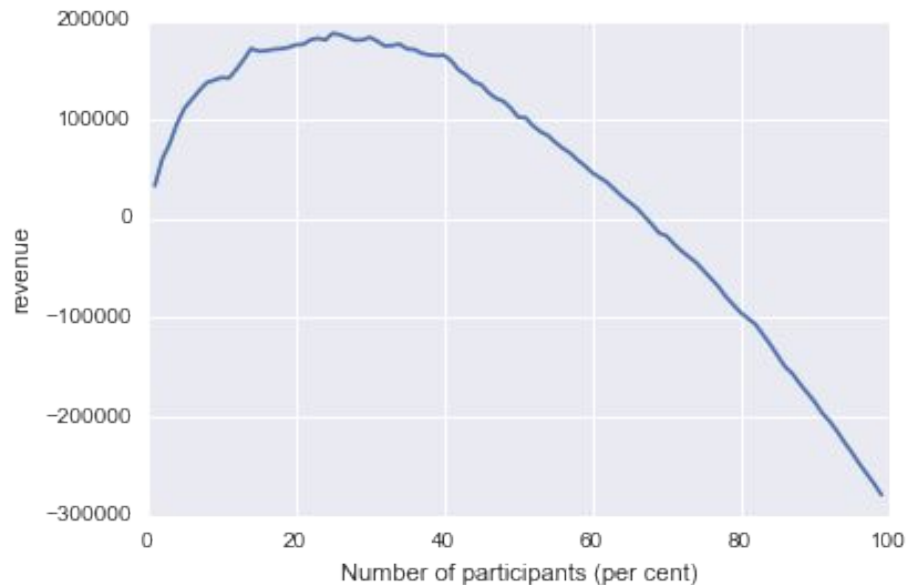
Если предположить, что средний пользователь ежемесячно приносит 2500 единиц, при этом если мы потратим 125 единиц на удержание, то можем получить следующую зависимость полученной прибыли к количеству участников кампании по удержанию.

Потенциальная прибыль при оптимальном подборе параметров на 10000 пользователей равна 238 750 единиц. Если в среднем телеком имеет 100млн. пользователей, то этот суммарный доход будет равен почти 1.9 млрд единиц.



# Экономическая модель: скидка на услуги

Также была рассмотрена другая модель, в которой пользователю, склонному к оттоку может быть предложена скидка на услуги компании. Аналогично, пользователь ежемесячно приносит 2500 единиц, скидка 5%. Зависимость полученной прибыли к количеству участников кампании по удержанию представлена на графике. Оптимальное значение прибыли несколько скромнее предыдущей модели: 187 312 единиц на 10000 пользователей компании.



# Советы по внедрению модели

1. Перед внедрением акции по удержанию пользователей, необходимо провести A/B тестирование, чтобы дополнительно оценить вероятность участия в акции пользователя, склонного к оттоку, и оценить экономический эффект, достигнутый с помощью удержания. Эксперимент может быть проведен в следующем виде: выделить 2 равных группы пользователей, и в первой группе пользователей в оптимальном топе пользователей провести кампанию по удержанию.
2. Модель может быть улучшена, если увеличится размер данных, а также количество объектов в выборке. Кроме того, если часть информации о некоторых признаках может быть раскрыта, можно провести дополнительную разработку признаков, что в итоге улучшит модель.
3. Кроме того, необходимо наблюдать за моделью и изучать ее поведение во времени и эффективность в предсказании оттока с использованием фактических данных.

# Выводы

Были проанализированы и обработаны данные, предоставленные телеком компанией Orange. С помощью проанализированных данных успешно построена и оценена модель, которая позволит в дальнейшем определять заблаговременно пользователей склонных к оттоку. Построены экономические модели для оценки эффективности кампании по удержанию пользователей. Описаны практические советы по использованию модели и внедрению ее в использование. Цель проекта успешно достигнута.