|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**  **«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ**  **(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»** | | | | |
| **Журнал практики** | | | | |
| Институт № 8 | «Компьютерные науки и прикладная математика» | | | |
|  |  | | | |
| Кафедра | \_\_\_\_804\_\_\_\_\_\_\_\_\_\_ | | Учебная группа | М8О-201М-21 |
|  |  | |  |  |
| ФИО обучающегося | | \_\_\_\_\_\_ Фейзуллин Кирилл Маратович \_\_\_\_\_\_\_\_\_\_\_\_ | | |
|  | |  | | |
| Направление подготовки/  специальность | | \_\_01.04.04 «Прикладная математика» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | |
|  | | *шифр, наименование направления подготовки/специальности* | | |
|  | |  | | |
| Вид практики | | \_Производственная (Профессиональная практика)\_\_\_\_\_\_\_\_ | | |
|  | | *учебная, производственная, преддипломная или другой вид практики* | | |
| Оценка за практику | | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Платонов Е. Н. | | |

Москва

2022

|  |  |  |
| --- | --- | --- |
| 1. **Место и сроки проведения практики:** | | |
| Наименование организации: | МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ (НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ) кафедра 804 | |
| Сроки проведения практики |  | |
| дата начала практики: | \_\_\_01 сентября 2022 г.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |
| дата окончания практики: | \_\_\_04 января 2023 г.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |
| 1. **Инструктаж по технике безопасности:** | | |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | /\_ Платонов Е.Н.\_/ | 01 сентября 2022 г. |
| *подпись проводившего* | *расшифровка подписи* | *дата проведения* |
| 1. **Индивидуальное задание обучающегося:** | | |
| Разработать модель Decision trees for uplift modeling. Смоделировать обучающую выборку на основе данных для UpLift моделирования от X5 Retail с добавлением характеристик заработка и списания бонусов. | | |
| 1. **План выполнения индивидуального задания обучающегося:** | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **№ п/п** | **Место проведения** | **Тема** | **Период выполнения** |
| 1 | Каф. 804 | Инструктаж. | 01.09.2022 |
| 2 | Каф. 804 | Знакомство с научной темой | 01.09.22-30.09.22 |
| 3 | Каф. 804 | Работа с библиографическими источниками и ресурсами Интернет по теме исследования | 01.10.22-31.10.22 |
| 4 | Каф. 804 | Разработка модели Decision trees for uplift modeling. | 01.11.22-20.12.22 |
| 5 | Каф. 804 | Оформление отчета. Подведение итогов. | 21.12.22-04.01.23 |

|  |  |  |
| --- | --- | --- |
| **Утверждаю** | | |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | /\_ Платонов Е.Н.\_\_\_\_/ | 01 сентября 2022 г. |
| *подпись руководителя от МАИ* | *расшифровка подписи* | *дата утверждения\** |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | /\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/ | 01 сентября 2022 г. |
| *подпись руководителя от организации/предприятия* | *расшифровка подписи* | *дата утверждения\** |
| **Ознакомлен** | | |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | /\_\_\_\_\_ Фейзуллин К.М.\_/ | 01 сентября 2022 г. |
| *подпись обучающегося* | *расшифровка подписи* | *дата ознакомления\** |

|  |  |  |
| --- | --- | --- |
| 1. **Отзыв руководителя практики от организации/предприятия:** | | |
| Фейзуллину К.М. была поставлена задача исследования и разработки методов UpLift моделирования для рекламной кампании. Во время исследования были разобраны различные подходы для моделирования прогнозируемого инкрементального отклика реципиентов от рассылки сообщений. Для исследования были реализованы методы моделирования с помощью одной модели, с помощью двух независимых моделей и с помощью регрессионной модели, на языке программирования Python,  Выявлен лучший метод в виде регрессионной модели и выявлена лучшая регрессионная модель – градиентный бустинг CatBoost, бесплатно распространяемый компанией Яндекс.  Поставленная задача решена в полном объёме и заслуживает оценки «отлично». | | |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | /\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/ | \_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_ 20\_\_г. |
| *подпись руководителя от организации/предприятии* | *расшифровка подписи* | *дата* |

1. **Отчет обучающегося по практике:**

Объектом исследования являются задача прогнозирования инкрементального отклика реципиентов рекламной кампании.

Цель работы – улучшение результатов работы UpLift моделирования методами машинного обучения.

## **Определение метрик для оценки качества UpLift моделирования**

Так как задача UpLift представляет собой задачу оценки (скор балл) эффекта от коммуникации на реципиента, то нет и истинных ответов. Получается, что не удастся использовать классические метрики, такие как Accuracy и PR AUC, основанные на матрице ошибок, для классификации или среднеквадратичная ошибка для задачи регрессии при трансформации классов.

### UpLift на первых k – процентах выборки

Самая простая и интуитивно понятная метрика, особенно для применения в бизнесе и для интерпретации.

Допустим, что на коммуникации в компании имеется скромный бюджет, который может обеспечить связь всего с 30% клиентской базы для побуждения к целевому действию. Тогда целью UpLift моделирования будет найти такой алгоритм, который лучше всех максимизирует эффект от коммуникаций на первых 30% клиентов.

Чтобы получить значение этой метрики, нужно ранжировать результат прогноза по убыванию, чтобы отобрать клиентов, на которых коммуникация оказывает наибольший эффект. Далее берется разница между конверсией целевой группы, с которой осуществлялась коммуникация, и конверсией контрольной группы, которая осталась без коммуникации.

Формула имеет следующий вид:

,

где .

Как и сам UpLift, имеет область значений [-1, 1].

Причем, данную метрику можно рассчитать двумя способами, в зависимости от ранжирования по прогнозу UpLift:

* Сортировка происходит по прогнозу и далее берется разность рабочей и контрольной группы.
* Сортировка происходит внутри каждой группы обособленно и далее берется разность.

Второй вариант имеет более практическое применение, так для оценки эффективности от коммуникаций при рекламных кампаниях, при планировании проведения мероприятий, образуются две однородные выборки – рабочая и тестовая группа.

Для дальнейшего исследования будем оценивать метрику при k = 30%.

### UpLift кривая (UpLift Curve)

Данная кривая строится как функция с нарастающим итогом, где для каждой точки задается соответствующий UpLift.

Определяется следующим образом:

, где

,

.

Аналогично и для контрольной группы.

Пример данной кривой на рисунке 2.



Рисунок

### Qini кривая

Данную функцию можно выразить через UpLift кривую следующим образом:

Данная кривая будет полезна в тех случаях, когда рабочая группа кратно превышает размер контрольной группы, с чем можно столкнуться во время исследования модели при внедрении в бизнес, когда у компании есть бюджет на произведение коммуникаций со всей клиентской базой и чтобы не упускать потенциальный доход, контрольная группа выделяется как можно меньше.

Таким образом будет получено инкрементальный эффект от коммуникаций в единицах измерения одного клиента.

## **Источник данных**

За источник данных были взяты результаты массовой рассылки СМС в ноябре на 473 861 человек. По истечении недели после рассылки появляется возможность определить целевую переменную (target): 0 – нет покупки в течении недели, 1 – есть покупка в течении недели. И так как нам известно заранее, кому была отослана СМС, а кому нет, очень просто определяется параметр коммуникации (treat): 0 – человек не получал СМС, 1 – человек получил СМС. Для клиентов из эксперимента были рассчитаны покупательские показатели за 4 месяца до момента рассылки, которые будут использованы как обучающие признаки.

Опишем набор данных детальнее. Он состоит из:

* Общая информации о клиентах и целевые переменные для обучения:

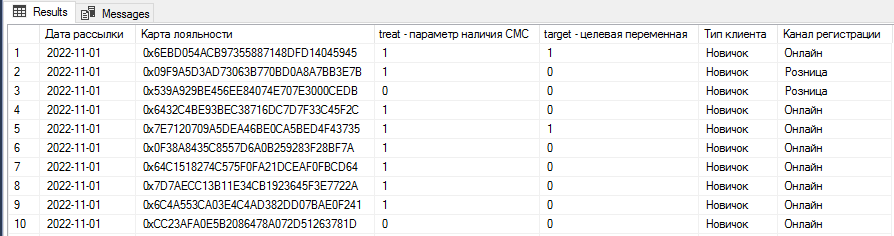


Рисунок 2

* История покупок клиентов до коммуникаций:

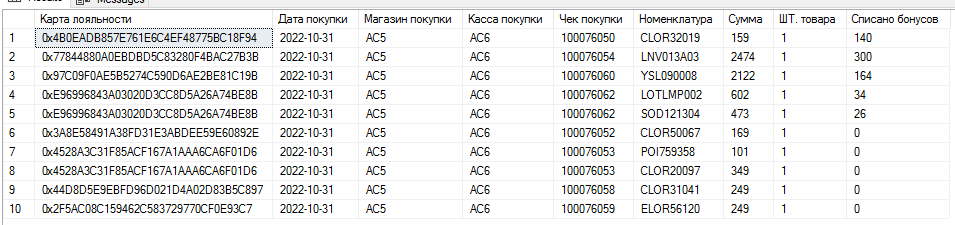


Рисунок 3

## **Анализ и агрегирование данных**

Так как данные для UpLift моделирования составляют находятся в базе SQL Server компании, то было решено и взаимодействовать с ними через реляционный язык запросов T-SQL. Для этого был использован менеджер запросов SQL Management Studio.

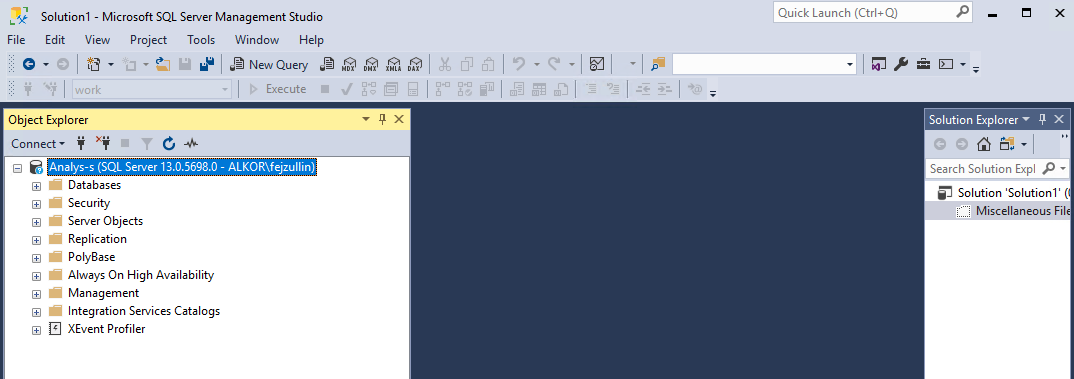


Рисунок 4

Для моделирования основных обучающих признаков был использован принцип RFM - сегментации[[1]](#footnote-1). То есть, по покупкам клиентов были определены следующие параметры:

* Частота покупок – количество покупок за расчетный период.
* Период с момента последней покупки.
* Сумма товарооборота с клиента за расчетный период - в нашем случае возьмем средний чек, так как это стратифицировать клиентов явным образом.

Также была собрана статистика по среднему времени между покупками, минимальном и максимальному интервалу между покупками, а также по трате и заработку бонусов программы лояльности, средняя скидка за счет бонусов, количество покупок и суммы с тратой всех бонусов, количество покупок и суммы с тратой заработанных бонусов, количество покупок и суммы с тратой начисленных в периоды акций бонусов. Вдобавок к этому были учтены и анкетные данные.

Таким образом было получено пространство из 32-ух обучающих признаков:

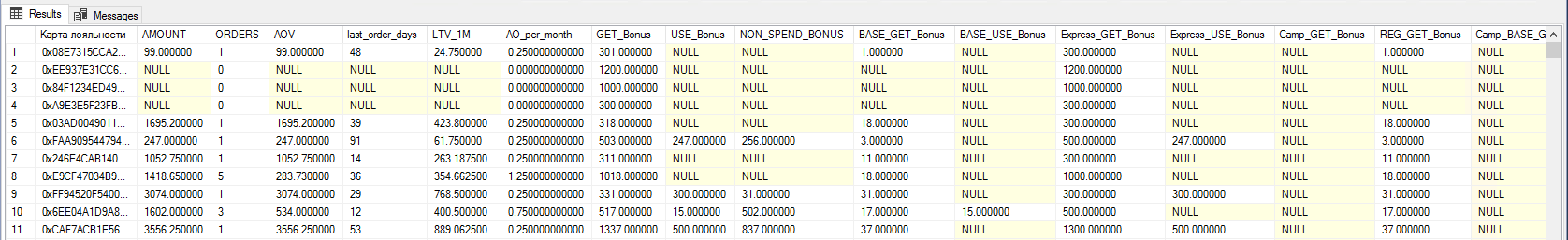


Рисунок 5

## **Реализация UpLift моделирования методами машинного обучения**

### Базовая модель

Перед проведением экспериментов следует определить базовую модель, от функционала качества которой нужно будет отталкиваться. Так как базовая модель предполагает слепое прогнозирование без обработки пространства признаков, в нашем случае подойдет равномерная случайная величина, распределенная от -1 до 1.

По итогам такого моделирования получаем следующие значения метрик:

* = 0.0073
* Qini curve AUC = -0.0016
* UpLift curve AUC = -0.0004

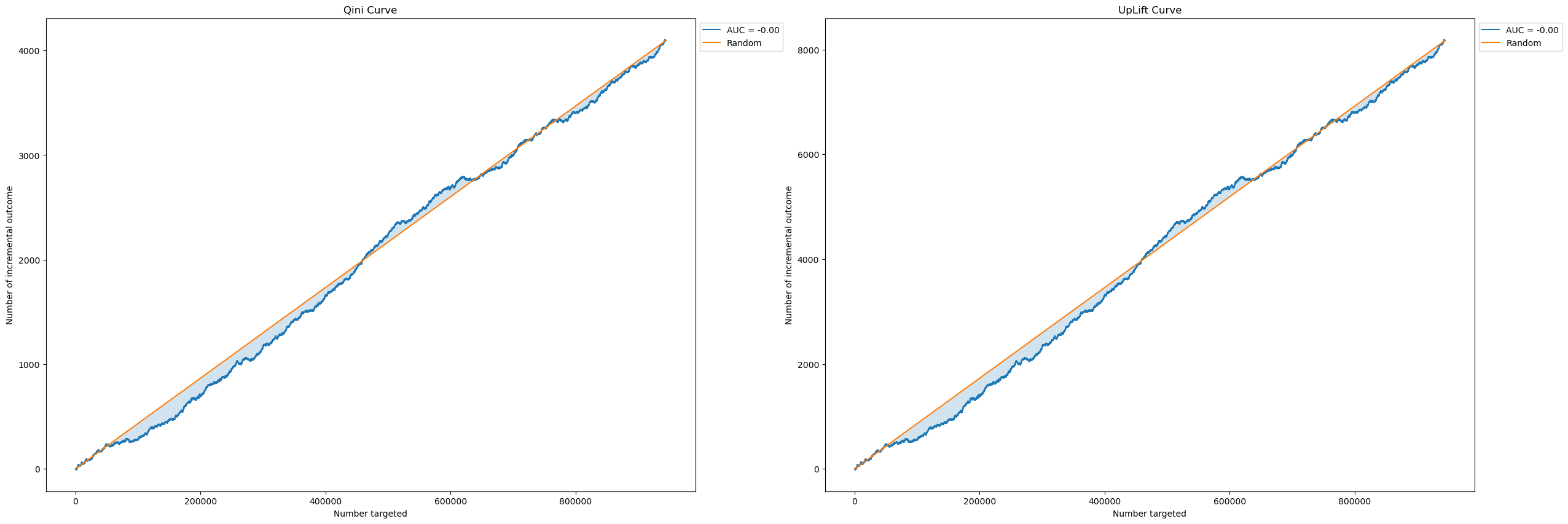


Рисунок 6. Графики кривой QINI и UpLift

### Экспериментальная установка

Исследование методов UpLift моделирования с помощью машинного обучения реализовано на высокоуровневом языке программирования Python, с использованием библиотек scikit-learn, scikit-uplift, CatBoost.

Для сравнения методов моделирования используется модель градиентного бустинга с базовыми параметрами, реализованный в библиотеке CatBoost.

Чтобы избежать ложных выводов по результатам работы модели на тестовом множестве, в исследовании используется кросс валидация c разбиением выборки на 5 долей. По итогу кросс валидации будет браться средняя по метрикам качества, на основе которых и будет сравнение. Иллюстрация работы кросс валидации на рисунке 13.



Рисунок 7. Схема кросс валидации

### Моделирование с одной моделью

Самое простое и понятное решение. На тренировочной выборке обучаем любую модель бинарной классификации по всем обучающим признакам, включая коммуникационную переменную.

Далее для тестовой выборки задаем коммуникационную переменную равную 1 и определяем прогноз вероятности, что объект совершит целевое действие.

Далее для тестовой выборки задаем коммуникационную переменную равную 0 и снова определяем прогноз вероятности, что объект совершит целевое действие.

После этого берется разность вероятностей при наличии коммуникации и при отсутствии, что и будет значением UpLift.

По итогам моделирования получены следующие усредненные метрики:

* = 0.0158
* Qini curve AUC = 0.0223
* UpLift curve AUC = 0.0055

По итогу кросс валидации имеются два типа событий:

* Когда uplift позволяет получить наибольший инкрементальный эффект, как на рисунке 8.

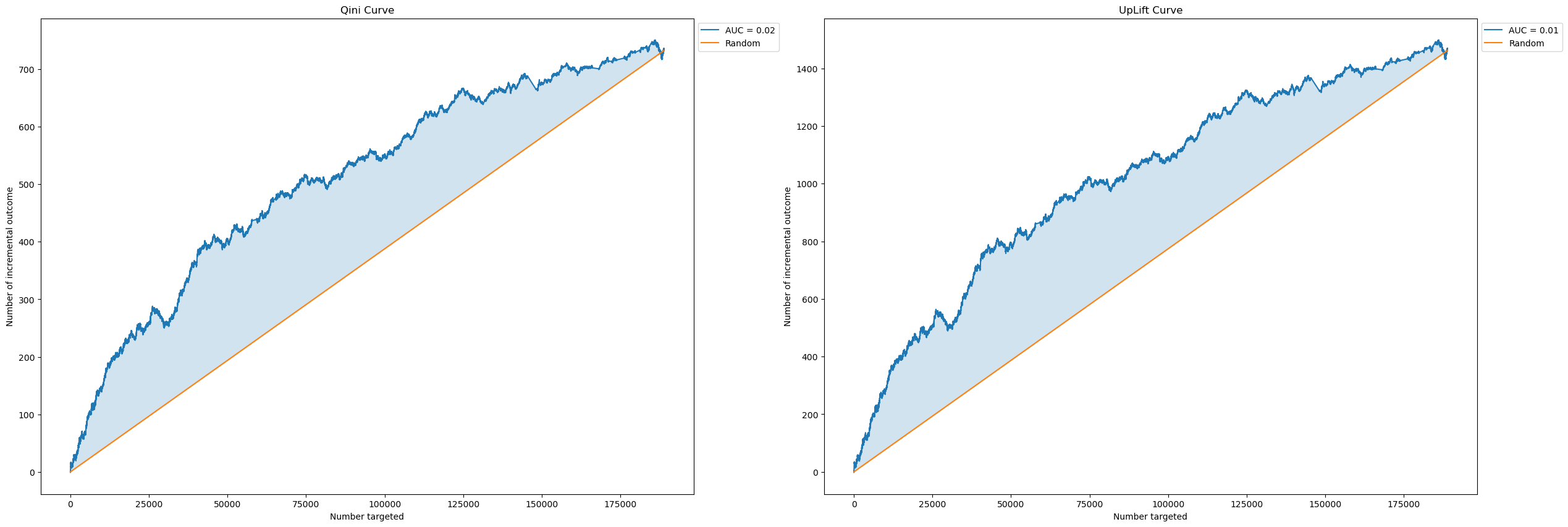


Рисунок 8.Графики кривой QINI и UpLift для результатов моделирования одной моделью в лучшем случае

* Когда uplift позволяет получить инкрементальный эффект с переменным успехом, как на рисунке 9.

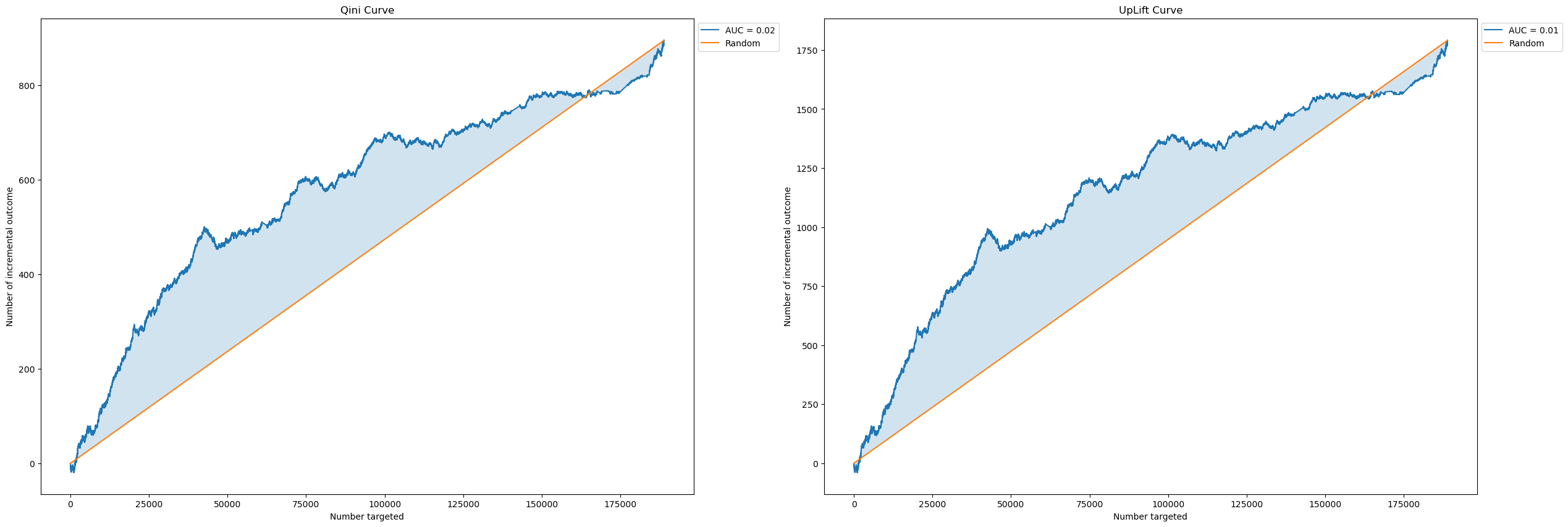


Рисунок 9. Графики кривой QINI и UpLift для результатов моделирования одной моделью в худшем случае

### Моделирование с двумя независимыми моделями

Метод представляет собой обучение двух независимых моделей на тренировочных данных, где одна модель обучается на целевой группе, а вторая обучается на контрольной. Далее на тестовых данных прогнозируется вероятность выполнения целевого действия для одной и для второй модели и берется их разность.

Но тут сразу возникает нюанс, что при отсутствии равного объема целевой и контрольной группы, модели не будут иметь одинаковую полноту обучения. Но в нашем случае этого происходить не будет, так как рабочая и тестовая группа равного объема.

По итогам моделирования получены следующие усредненные метрики:

* = 0.0144
* Qini curve AUC = 0.0167
* UpLift curve AUC = 0.0042

По итогу кросс валидации имеются два типа событий:

* Когда uplift позволяет получить наибольший инкрементальный эффект, как на рисунке 10.

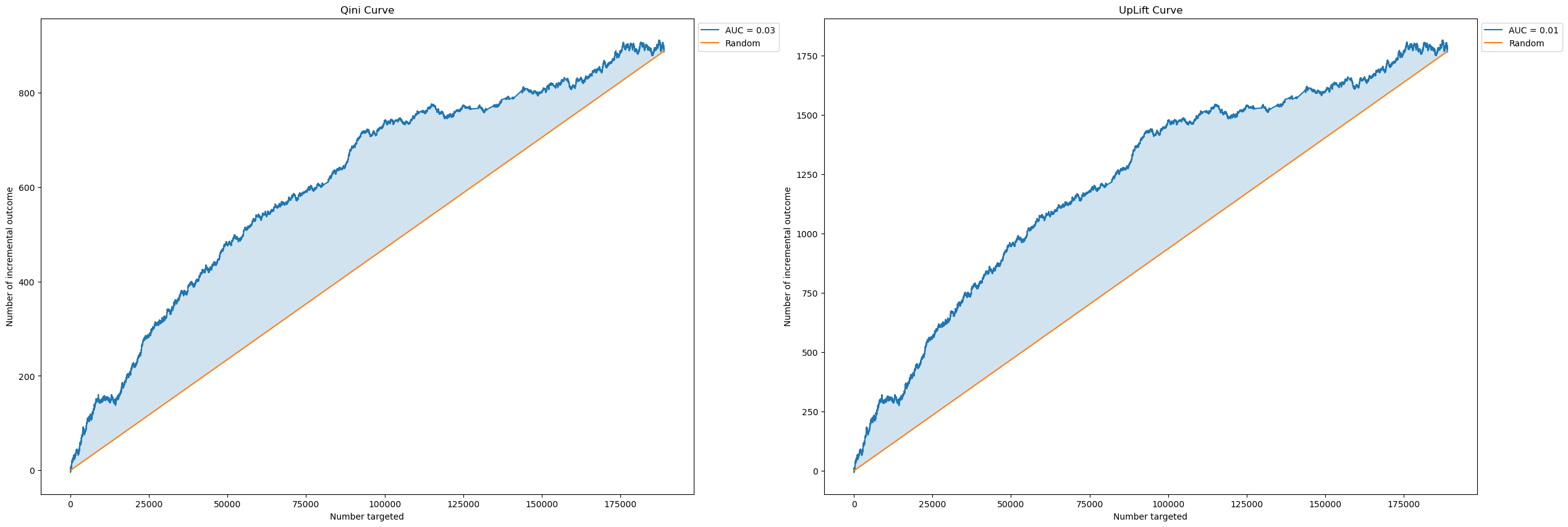


Рисунок 10. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в лучшем случае

* Когда uplift позволяет получить инкрементальный эффект с переменным успехом, как на рисунке 11.

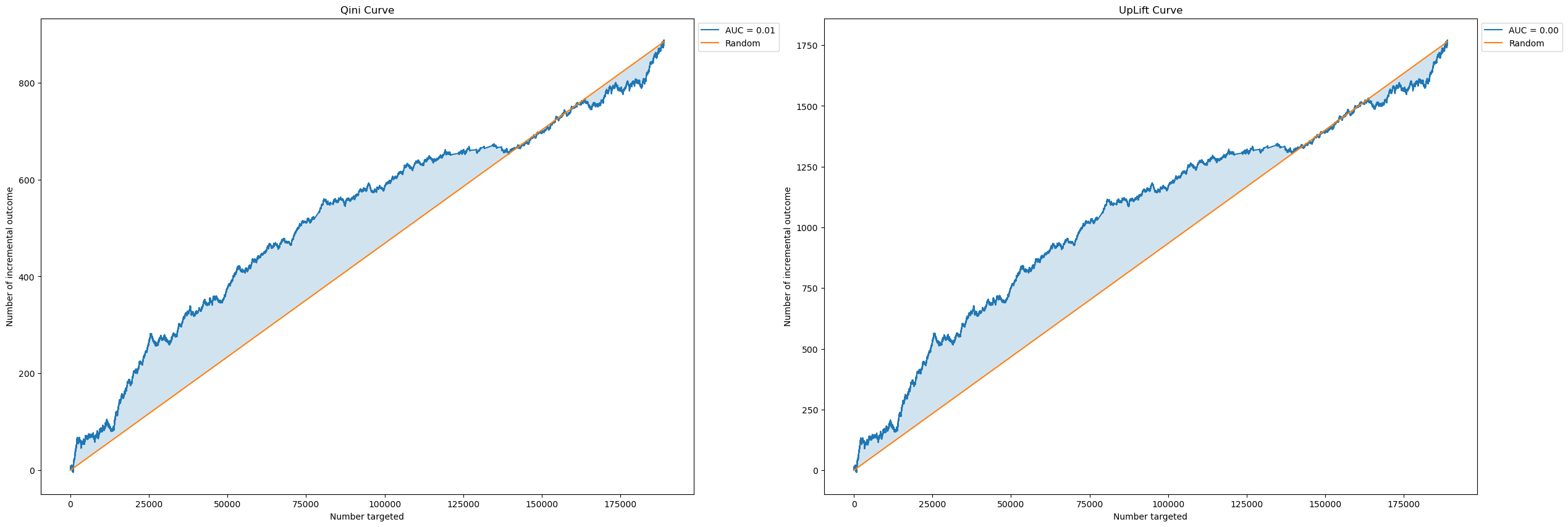


Рисунок 11. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в худшем случае

Так же стоит добавить, что поведение показателей качества обучения на тестовой выборке в 4 из 5 итераций кросс валидации выглядит как на рисунке 11, что говорит об ухудшении качества обучения – о чем и сигнализируют усредненные показатели , Qini curve AUC, UpLift curve AUC.

### Метод трансформации класса (задача классификации)

В данном методе мы вернемся снова к единой модели, но теперь преобразуем коммуникационную переменную и целевую переменную в одну следующим образом:

, где

Тогда трансформированный класс будет иметь следующие значения:

Далее произведем переход к задаче классификации для однозначной интерпретации прогноза.

По результатам моделирования были получены следующие усредненные результаты:

* = 0.0124
* Qini curve AUC = 0.0081
* UpLift curve AUC = 0.0022

По итогу кросс валидации имеются два типа событий:

* Когда uplift позволяет получить наибольший инкрементальный эффект, как на рисунке 12.

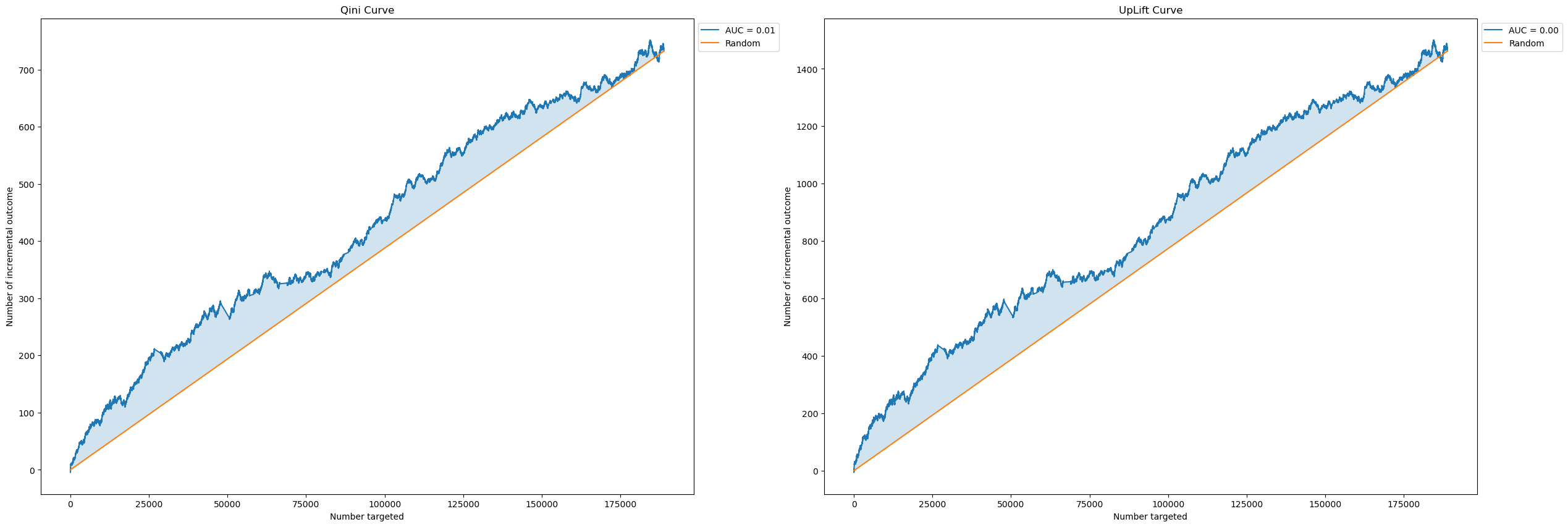


Рисунок 12. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в лучшем случае

* Когда uplift позволяет получить инкрементальный эффект с переменным успехом, как на рисунке 13.

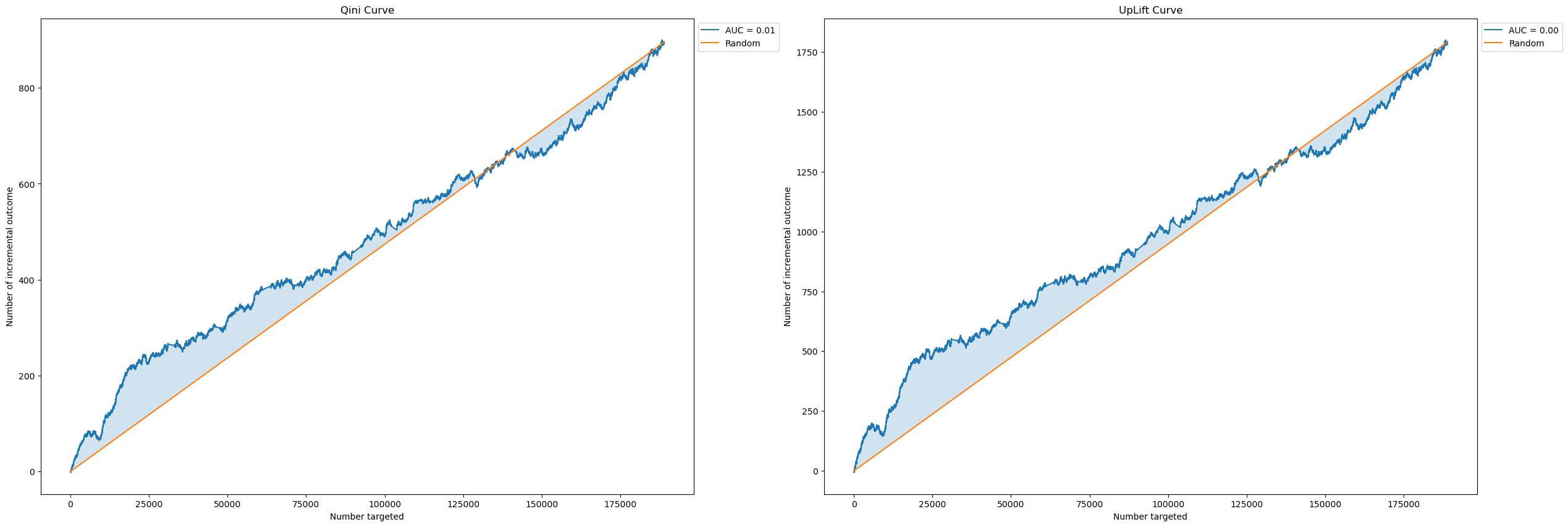


Рисунок 13. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в худшем случае

Метод трансформации класса в задаче классификации показывает еще более худшие показатели качества обучения, чуть ли не в 2 раза хуже, чем в моделировании двумя независимыми моделями.

### Метод трансформации класса (задача регрессии)

В данном методе мы вернемся снова к единой модели, но теперь преобразуем коммуникационную переменную и целевую переменную в одну следующим образом:

, где

.

В нашем случае, . Тогда трансформированный класс будет иметь следующие значения:

Далее произведем переход к задаче регрессии для однозначной интерпретации прогноза.

По результатам моделирования были получены следующие усредненные результаты:

* = 0.0138
* Qini curve AUC = 0.0155
* UpLift curve AUC = 0.0038

По итогу кросс валидации имеются два типа событий:

* Когда uplift позволяет получить наибольший инкрементальный эффект, как на рисунке 14.

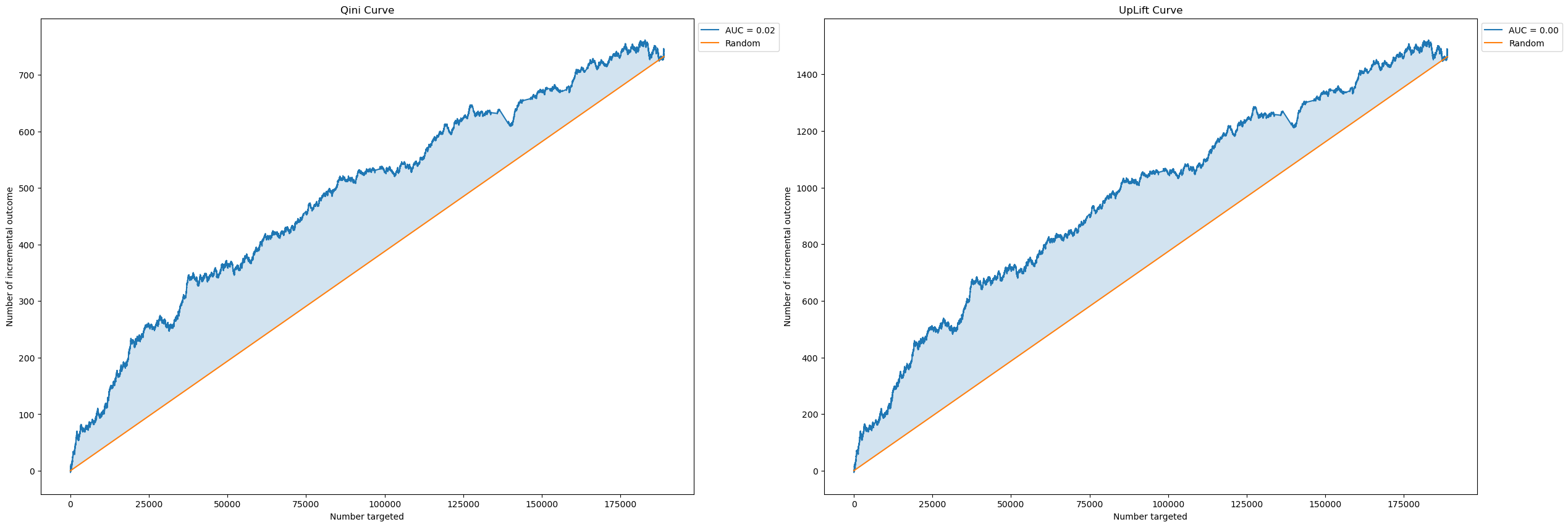


Рисунок 14. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в лучшем случае

* Когда uplift позволяет получить инкрементальный эффект с переменным успехом, как на рисунке 15.

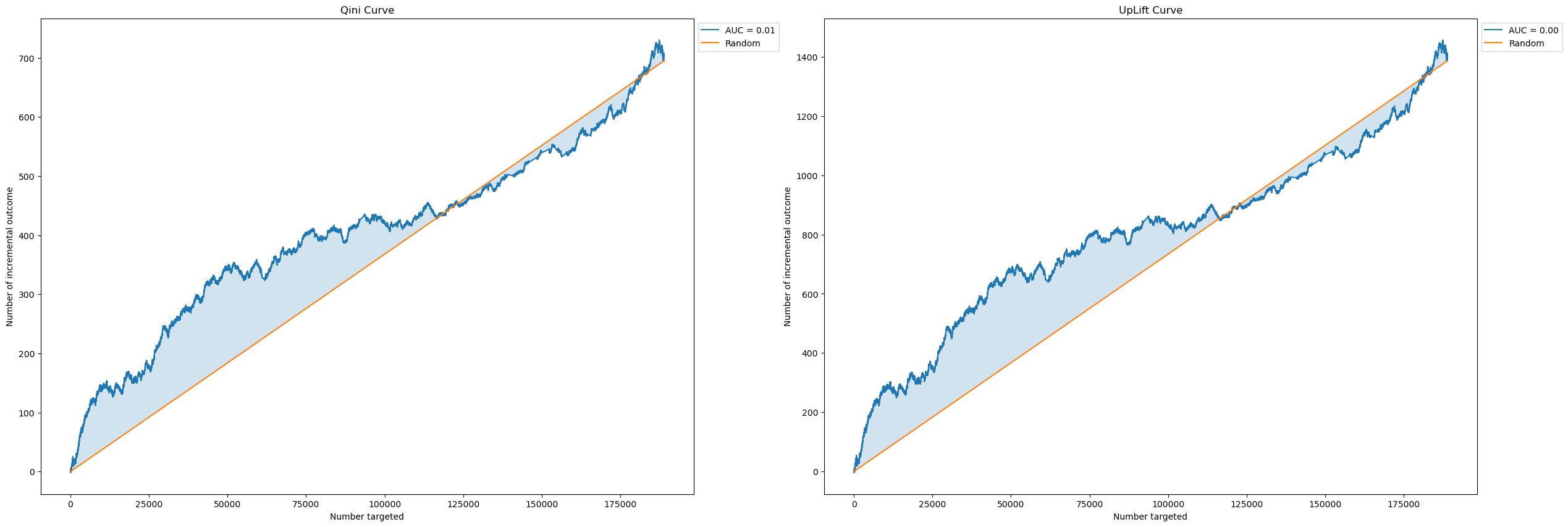


Рисунок 15. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в худшем случае

Данный метод чуть хуже, чем метод с двумя независимыми моделями.

### Исследований архитектур моделей машинного обучения

#### Поиск лучшей архитектуры для задачи классификации

Так как UpLift моделирование напрямую зависит от качества обучения на наших данных, чтобы максимизировать наши результаты, найдем наилучшую структуру модели классификации и найдем для нее целевые показатели.

Сравнение структур моделей будет происходить с помощью библиотеки evalml, которая содержит внутри себя уже весь реализованный функционал.

По итогам поиска по 13-ти моделей, наилучшие показатели имеет уже использованный ранее градиентный бустинг из библиотеки Яндекс CatBoost. Лучшие результаты в таблице 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Номер** | **pipeline\_name** | **validation\_score** | **percent\_better\_baseline** |
| 1 | Stacked Ensemble Classification Pipeline | 0,415 | 4047% |
| 2 | Random Forest Classifier w/ Label Encoder + Replace Nullable Types Transformer + Imputer + Undersampler | 0,415 | 4046% |
| 3 | LightGBM Classifier w/ Label Encoder + Replace Nullable Types Transformer + Imputer + Undersampler + Select Columns Transformer | 0,406 | 3965% |

Таблица

Далее взяли лучший PipeLine – ансамбль из моделей: Логистическая Регрессия, Случайный Лес, Дерево Решений, Градиентный бустинг LigthGBM, Расширенные Деревья (Extra Trees), Градиентный бустинг CatBoost, Градиентный бустинг XGBoost. И модель классификации, обрабатывающая результаты ансамбла – ElasticNet.

По результатам моделирования были получены следующие усредненные результаты:

* = 0.0233
* Qini curve AUC = 0.0543
* UpLift curve AUC = 0.0136

По итогу кросс валидации имеются два типа событий:

* Когда uplift позволяет получить наибольший инкрементальный эффект, как на рисунке 16.

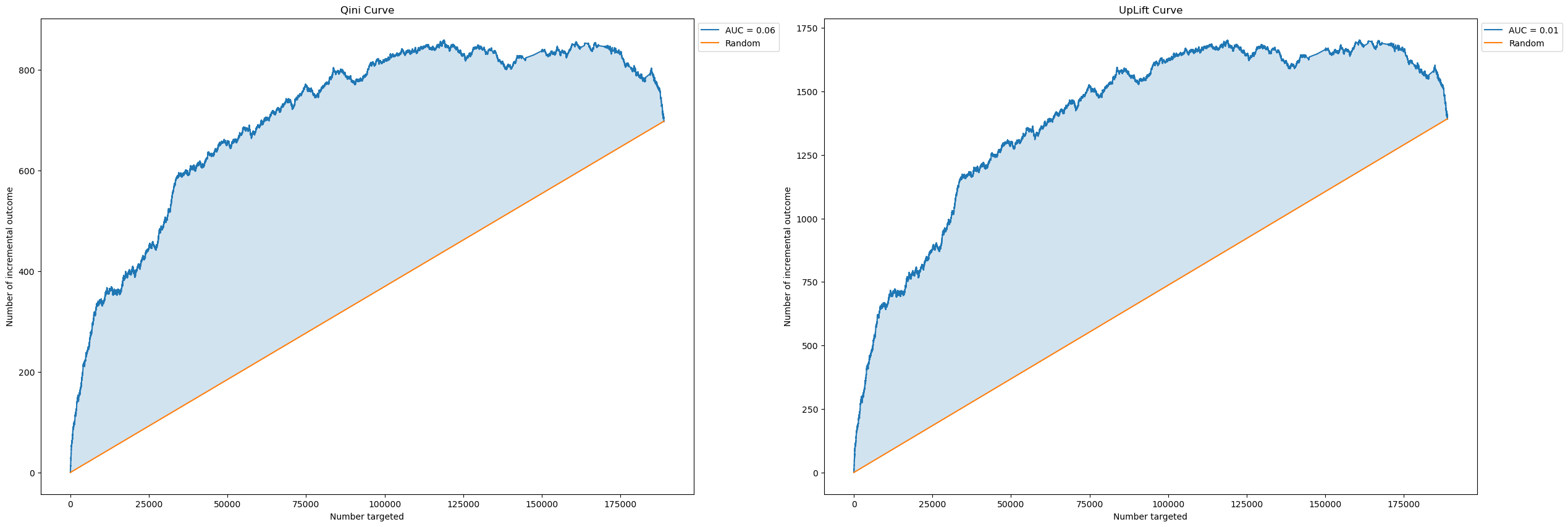


Рисунок 16. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в лучшем случае

* Когда uplift позволяет получить инкрементальный эффект с переменным успехом, как на рисунке 17.

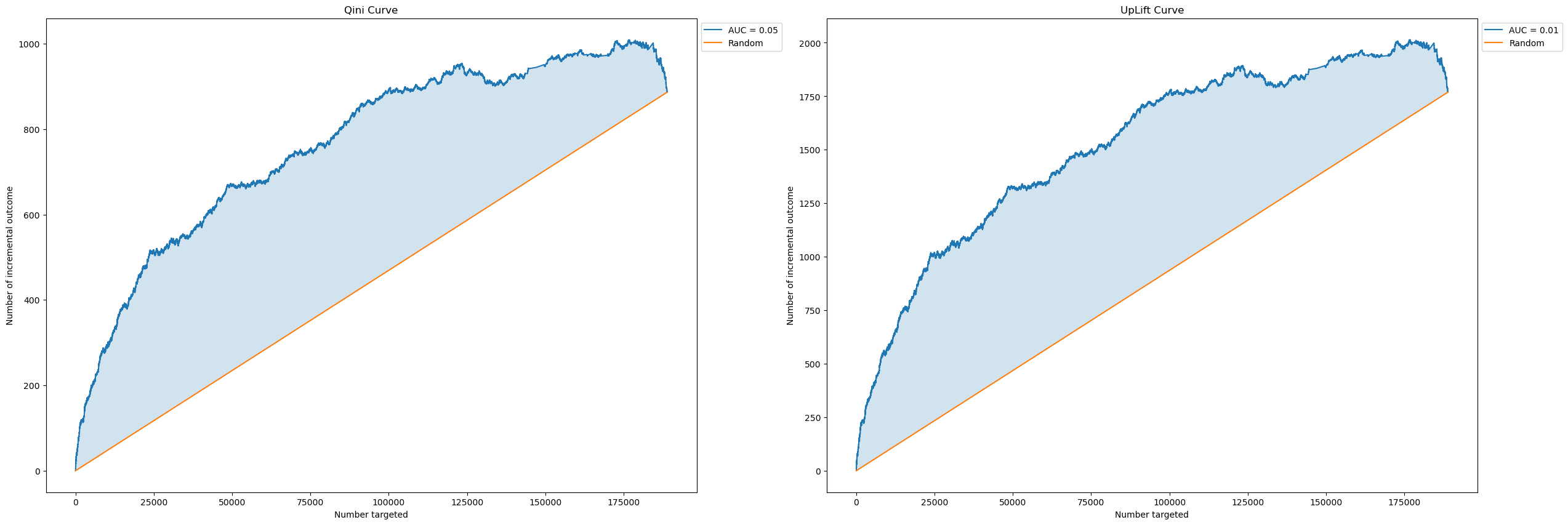


Рисунок 17. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в худшем случае

#### Поиск лучшей архитектуры для задачи регрессии

Так как по результатам подходов наилучшие имеет метод трансформации классов с переходом к задаче регрессии, то возникает вопрос – какая модель позволяет получить наилучший результат для нашей задачи.

Если считать, что наши целевые переменные достоверные, то косвенно оценивать качество моделей для сравнения можно и с помощью среднеквадратичной ошибки. Ведь та модель, которая лучше всего обучиться на тренировочных данных и тестовых данных и должна потенциально иметь наилучший UpLift на практике.

Сравнение структур моделей будет происходить с помощью библиотеки evalml, которая содержит внутри себя уже весь реализованный функционал.

По итогам поиска по 11-ти моделям, наилучшие показатели имеет уже использованный ранее градиентный бустинг из библиотеки Яндекс CatBoost. Лучшие результаты в таблице 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Номер** | **pipeline\_name** | **validation\_score** | **percent\_better\_baseline** |
| 1 | CatBoost Regressor w/ Replace Nullable Types Transformer + Imputer + Select Columns Transformer | 0,27092 | 0,3873% |
| 2 | Elastic Net Regressor w/ Replace Nullable Types Transformer + Imputer + Standard Scaler + RF Regressor Select From Model | 0,27093 | 0,2225% |
| 3 | Mean Baseline Regression Pipeline | 0,27093 | 0,0000% |

Таблица

Далее взяли лучший PipeLine: регрессионная модель градиентного бустинга от Яндекс - CatBoost, с выбором наиболее значимых для модели параметров.

По результатам моделирования были получены следующие усредненные результаты:

* = 0.0179
* Qini curve AUC = 0.0314
* UpLift curve AUC = 0.0077

По итогу кросс валидации имеются два типа событий:

* Когда uplift позволяет получить наибольший инкрементальный эффект, как на рисунке 18.

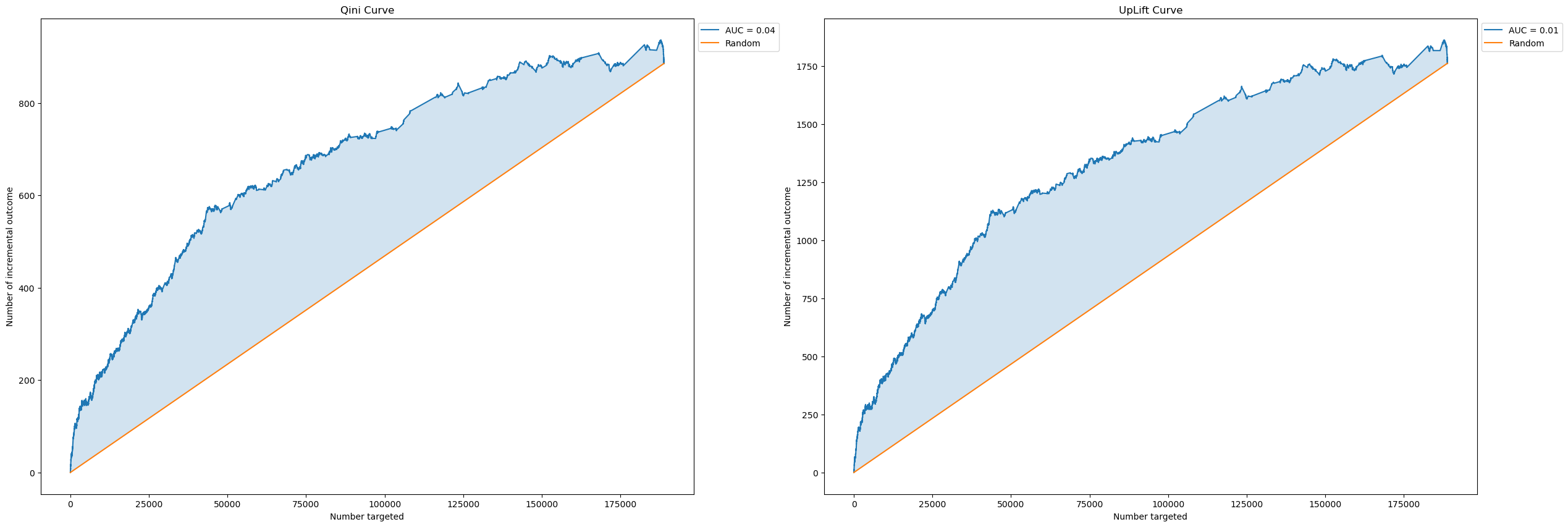


Рисунок 18. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в лучшем случае

* Когда uplift позволяет получить инкрементальный эффект с переменным успехом, как на рисунке 19.

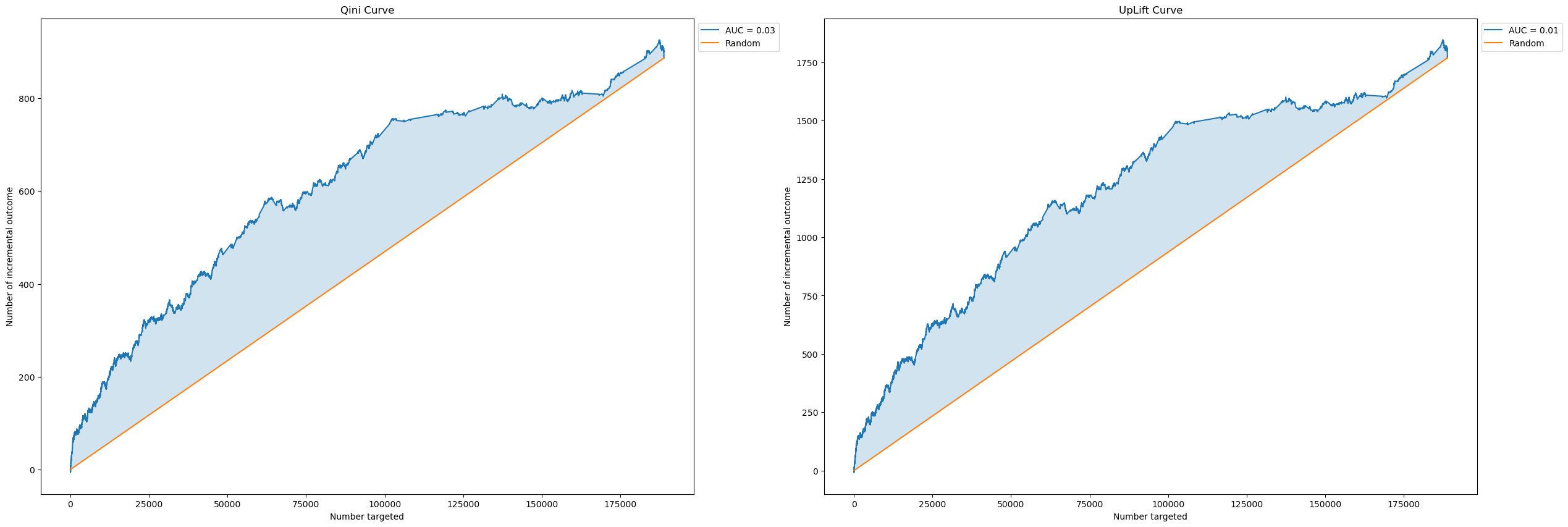


Рисунок 19. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в худшем случае

### Сравнительные результаты

## **Заключение**

В данной работе были исследованы методы моделирования UpLift с помощью машинного обучения на исходных данных от X5 Retail Group, выложенных в открытый доступ.

В работе были рассмотрены метрики оценивания качества прогноза UpLift при алгоритме с одной моделью, при алгоритме с двумя независимыми моделями и при работе с одной моделью после трансформации классов и перехода к задаче регресии.

По итогам моделирования с данными обучающими признаками, лучшее качество имеет метод трансформации классов.

После определения метода было решено найти наилучшую структуру модели с помощью AutoML конвейров. В результате чего выяснилось, что с данными признаками лучшей моделью является градиентный бустинг в библиотеке CatBoost от компании Яндекс.

Причем, данная модель имеет наилучшие показатели по всем целевым метрикам по итогам усреднения результатов кросс – валидации.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Источник данных // [https://ods.ai/competitions/x5-retailhero-uplift-modeling](%20https:/ods.ai/competitions/x5-retailhero-uplift-modeling)

2. RF – сегментация // <https://www.moengage.com/blog/rfm-analysis-using-rfm-segments/>

|  |  |  |
| --- | --- | --- |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | /\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/ | \_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_ 20\_\_г. |
| *подпись обучающегося* | *расшифровка подписи* | *дата* |

1. RFМ – сегментация // <https://www.moengage.com/blog/rfm-analysis-using-rfm-segments/> [↑](#footnote-ref-1)