ПРОГНОЗИРОВАНИЕ СКЛОННОСТИ ПОКУПКИ КЛИЕНТОВ НА ПРИМЕРЕ КОМПАНИИ «SIBERIAN WELLNESS»

А. И. Душенин

Новосибирский государственный университет, Россия
Email: a.dushenin@g.nsu.ru

В работе рассматривается задача прогнозирования вероятности покупки клиентов в следующем месяце на примере компании «Siberian Wellness» (место работы автора). Определены признаки, влияющие на вероятность покупки клиентов. Представлена методология построения моделей прогнозирования вероятности покупки клиентов. Построены соответствующие модели бинарной классификации для четырёх стран: Россия, Турция, Узбекистан, Вьетнам. Рассчитаны метрики качества моделей на независимых данных.

PREDICTION OF CUSTOMERS' PURCHASE PROSPENSITY ON THE EXAMPLE OF THE COMPANY "SIBERIAN WELLNESS"

A. I. Dushenin

The paper considers the problem of predicting the probability of buying customers in the next month using the example of the Siberian Wellness company (the author's place of work). Signs that affect the likelihood of customers buying are identified. The methodology for constructing models for predicting the probability of customers buying is presented. The corresponding binary classification models are built for four countries: Russia, Turkey, Uzbekistan, Vietnam. Model metrics are measured on independent data.

Введение. Сегодня огромное число компаний используют инструменты анализа данных для принятия бизнес-решений и разработки стратегий на будущий период. Примерами таких подходов является RFM-сегментация и A/B-тестирование, проводимые с целью определения эффективного способа воздействия на клиентов. Однако для некоторых кейсов инструментов статистического анализа недостаточно, что побуждает использовать модели машинного обучения для оптимизации деятельности компаний. Одной из таких задач является определение вероятности покупки клиентов в следующем месяце (периоде) по их предыстории покупок. **Цель работы** – построить модель прогнозирования склонности покупки клиентов для компании «Siberian Wellness» для 4-х стран (Россия, Турция, Узбекистан, Вьетнам).

Значения вероятностей покупки клиентов позволяют определять индивидов, на которых необходимо оказывать мотивирующее воздействие определённой степени. Так, чем меньше вероятность покупки, тем сильнее нужно простимулировать клиента для покупки. Кроме того, если перейти от вероятностям к классам, то путём их агрегации можно оценить общее количество активных клиентов (или количество клиентов в определённом сегменте), что помогает

компании в стратегическом планировании (например, планирование бюджета на будущее).

Целевой переменной модели является факт покупки клиентом любого товара (1, если купил, 0 – иначе). Для построения модели (и подбора параметров) используются данные с 01.07.2020 по 30.06.2021 (строится прогноз на июль 2021 года). Для тестирования модели используются данные с 01.08.2020 по 31.07.2022 (строятся прогнозы на август 2021 – август 2022 гг.). Таким образом, обучающая и тестовая выборки не пересекаются.

Этап 1 – Генерация признаков. Оценка склонности клиентов к покупке побуждает исследование их паттернов поведения. Поэтому для построения модели используется информация о покупках клиентов за последние 12 месяцев (такое временное окно выбрано в связи с особенностями сегментации клиентской базы компании «Siberian Wellness»), а именно суммы покупок, количество транзакций, количество уникальных транзакций по месяцам. Для учёта товарной структуры для каждого месяца добавлены суммы покупок по каждой товарной категории. Для учёта сегментации для каждого месяца добавлены дамми-переменные, показывающие, был ли клиент уже зарегистрирован в этом месяце. К примеру, если индивид является клиентом компании полгода, то для него дамми-переменные принимают значения 0 для первых 6 месяцев и 1 для последних 6 месяцев временного окна. Помимо того, что эти факторы отражают «срок жизни» клиента, они уточняют вышеперечисленные признаки: если сумма покупок равна нулю, а дамми-переменная равна единице, то клиент не покупал товары в этом месяце (если равна нулю, то клиент не был зарегистрирован в этом месяце). Также учитывается город доставки клиента (категориальная переменная).

Этап 2 – Отбор признаков. Перед разработкой основной модели построена промежуточная модель для отбора признаков (для сокращения переобучаемости и увеличения скорости работы основного ML-алгоритма) – градиентный бустинг из библиотеки lightgbm. Для обработки категориальной переменной города использован CatBoostEncoder. В рамках исследования признак считается важным, если его значение feature_importance больше нуля. В итоге переменные, обозначающие суммы покупок по товарным категориям, не оказались важными. Удаление соответствующих факторов значительно сократило размерность признакового пространства. В табл. 1 представлен итоговый список факторов.

Этап 3 – Подбор гиперпараметров. В качестве основного ML-алгоритма выбран градиентный бустинг из библиотеки lightgbm (для учёта нелинейных связей). Для обработки категориальной переменной города использован CatBoostEncoder. Подбор гиперпараметров производился с помощью кросс-валидации обучающей выборки путём перебора по сетке. При этом максимизировалась метрика ROC AUC, т.к. неизвестны пороговые значения вероятностей.

Таблица 1

Итоговый список признаков для каждого клиента (49 признаков)

Признак	***Описание***
sum_1,…, sum_12	Суммы покупок по месяцам
count_1,…, count_12	Количество транзакций по месяцам
uniq_1,…, uniq_12	Количество уникальных транзакций по месяцам
life_1,…, life_12	Дамми-переменные «жизни» клиента по месяцам
city	Город доставки

Этап 4 – Подбор порогового значения. При определении класса модель использует следующее правило: если прогнозная вероятность больше порогового значения (threshold), то модель присваивает такому наблюдению положительный класс, в противном случае, ноль. Подбор порогового значения производился на обучающей выборке путём максимизации метрики F1. В табл. 2 представлены результаты подбора для каждой из стран.

Таблица 2

Пороговые значения для 4-х стран

Страна	***Значение***
Россия	0.33
Турция	0.35
Узбекистан	0.38
Вьетнам	0.28

Этап 5 – Тестирование модели. Для оценки качества моделей склонности к покупке рассматривались такие метрики, как ROC AUC (зелёная линия), PR AUC (красная линия), Precision (синяя линяя), Recall (чёрная линия), F1 (оранжевая линия) и Accuracy (серая линия). На рис. 1 представлены результаты.

По графикам можно сказать, что модели имеют приемлемое качество, как абсолютное (высокие значения метрик), так и относительное (относительно некоторого baseline). К примеру, отношение PR AUC к доле целевого события показывает, во сколько раз модель лучше случайного классификатора. Стоит отметить, что модели являются стабильными во времени (если не считать Турцию, для которой в последние месяцы начался спад уровня целевого события).

Также можно заметить, что значения Precision и Recall для некоторых стран имеют незначительные различия в динамике. Это соответствует тому, что число реальных положительных и число прогнозных положительных классов примерно совпадают. Такую тенденцию моделей можно использовать для прогнозирования количества активных клиентов (путём агрегации клиентских прогнозов). На рис. 2 представлены результаты сравнения реальной (зелёная линия) и прогнозной (красная линия) активной клиентской базы.

Заключение. Таким образом, в работе продемонстрирован подход для оценки склонности покупки клиентов на примере реальной компании. В дальнейшем предполагается расширение горизонта прогнозирования (например, на два/три месяца вперёд).

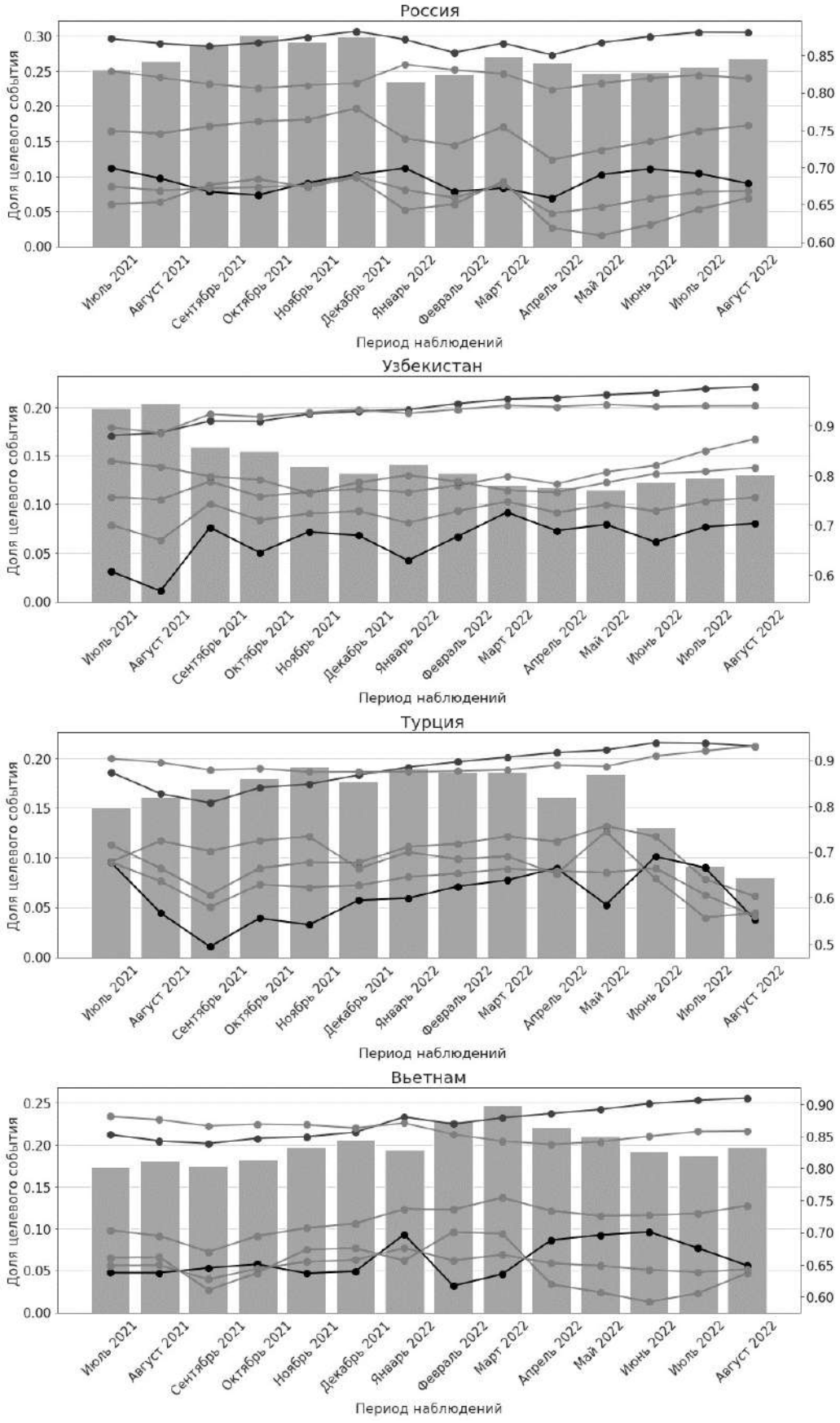

Рис. 1. Метрики качества прогнозов склонности к покупке

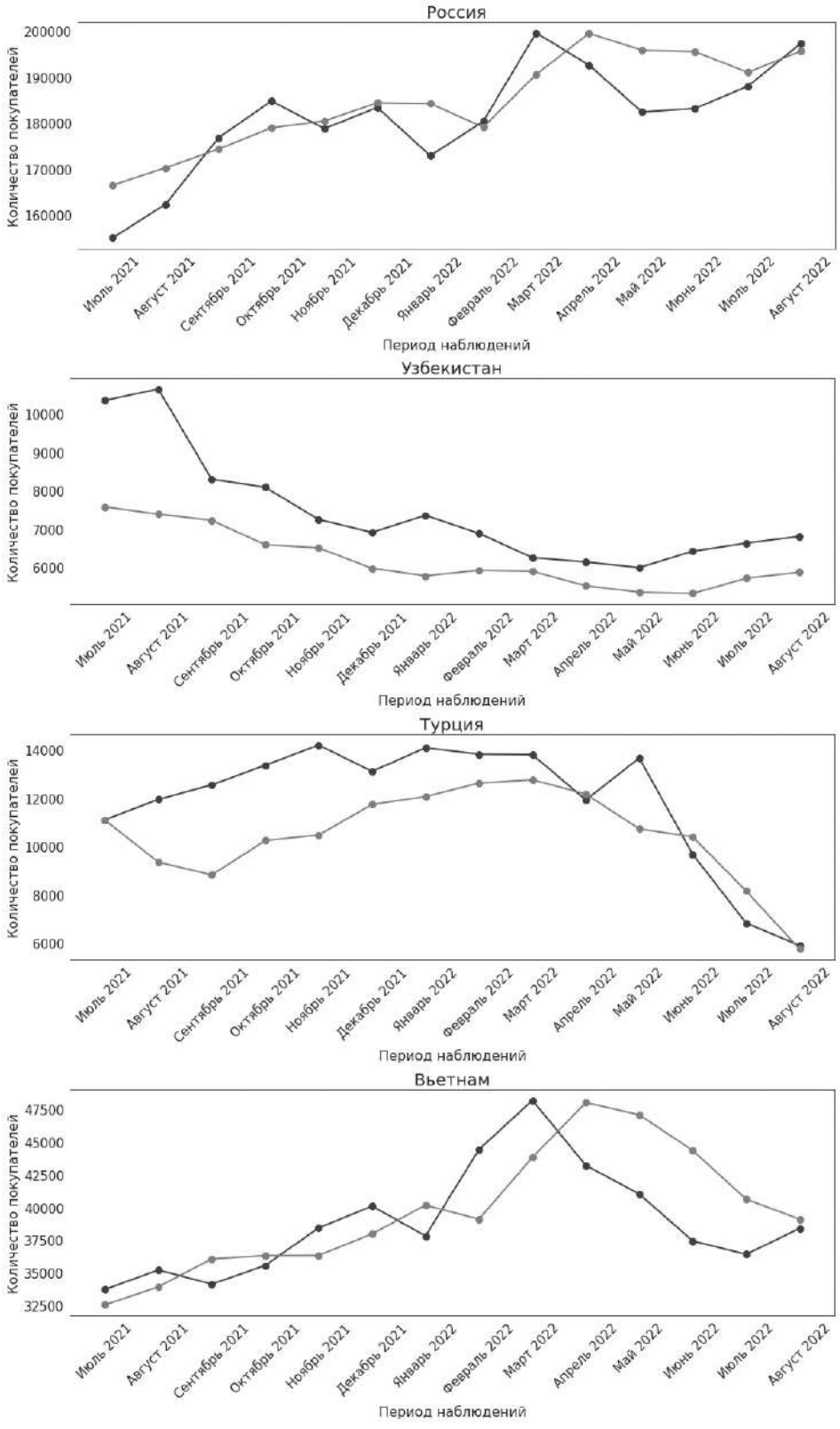

Рис. 2. Сопоставление количества реальных
и прогнозных покупателей