Precision-Recall \ AutoML \ метрика Gini

MPP-challenge: моделирование прогноза качества модели

На весенней конференции Scoring Day[1] команда из ВТБ рассказала про MPP-подход, с помощью которого можно строить модели предсказания качества других моделей. В своем докладе Денис Суржко, возглавляющий Data Science подразделение ВТБ, предложил банкам начать использовать MPP-подход и исследовать ряд важных вопросов о его применимости. Мы в «Ренессанс Кредите» принимаем этот челлендж и в этой статье расскажем о нашем опыте внедрения MPP-подхода, а также постараемся ответить на озвученные командой ВТБ вопросы.

**Сергей АФАНАСЬЕВ**, *КБ «Ренессанс Кредит», исполнительный директор, начальник управления статистического анализа*
**Диана КОТЕРЕВА**, *КБ «Ренессанс Кредит», руководитель направления моделирования и оперативного анализа*
**Константин СТАРОДУБ**, *КБ «Ренессанс Кредит», ведущий аналитик направления моделирования и оперативного анализа*

# MPP-challenge: моделирование прогноза качества модели

## Проблемы классических методов мониторинга моделей

В процессе эксплуатации моделей возникает необходимость в мониторинге их качества. Качество моделей можно оценивать с помощью различных метрик, среди которых банковским стандартом для оценки качества скоринговых карт является Gini. При мониторинге рисковых скоринговых карт возникает проблема «вызревания» целевой переменной, когда просрочка, по которой рассчитывается Gini, может «созревать» от нескольких месяцев до года. Таким образом, при мониторинге рисковых скоркарт метрика Gini является «запаздывающей» и не отражает текущего состояния работы скоркарт. Чтобы решить эту проблему, на практике используется подход мониторинга входящих в модель переменных (или других факторов), данные по которым можно получить в момент принятия решения.

Для мониторинга изменения характеристик входящего потока по переменным сравниваются их распределения, полученные в разные

[1] scorconf.ru/online_meetup_14_05.

даты срезов (например, ежедневно, еженедельно, ежемесячно и т.д.). Для сравнения двух распределений можно использовать различные метрики, среди которых общепринятой для мониторинга является показатель PSI (Population Stability Index), оценивающий изменение распределений во времени.

Мониторинг качества модели с помощью показателя PSI имеет ряд недостатков:

1) PSI отслеживает изменения переменных модели, что не обязательно отражает изменения в качестве предсказательной способности модели;

2) мониторинг PSI по каждой отдельной переменной отслеживает однофакторные изменения, то есть не учитывает природу многомерных зависимостей;

3) при большом количестве переменных в модели не всегда понятно, какой вклад в изменение качества модели вносит та или иная переменная.

MPP-модель позволяет строить прогнозы по выбранной метрике качества исходной модели в момент ее работы, то есть не дожидаясь вызревания оцениваемой метрики.

Проблему п. 3 на практике можно решить путем расчета средневзвешенного PSI по всем переменным, где в качестве весов можно взять абсолютные значения коэффициентов регрессии (если модель была построена с помощью регрессионных методов) или важностей переменных (для алгоритмов Random Forest, LightGBM и др.). Однако все остальные недостатки PSI остаются актуальными, то есть мониторинг показателя PSI по переменным не позволяет однозначно ответить на вопрос: изменится ли качество модели при изменении PSI переменных и насколько значимым окажется это изменение?

## MPP-подход

Избавиться от перечисленных недостатков классических методов мониторинга помогает подход Model Performance Predictor (MPP). В рамках MPP-подхода строятся две модели:

1) исходная модель, задача которой — предсказывать целевую переменную (например, факт наступления дефолта);

2) MPP-модель (строится на переменных исходной модели), цель которой — предсказывать выбранную метрику качества исходной модели.

Таким образом, MPP-модель позволяет строить прогнозы по выбранной метрике качества исходной модели в момент ее работы, то есть не дожидаясь вызревания оцениваемой метрики[1].

[1] Более подробно про MPP-подход можно почитать в статье команды ВТБ и в оригинальной научной статье: Кулик В., Коновалихин М., Суржко Д. Эволюция системы мониторинга моделей: нелинейный подход // Риск-менеджмент в кредитной организации. 2020. № 2; Ghanta S. et al. {MPP}: Model Performance Predictor. 2019 {USENIX} Conference on Operational Machine Learning (OpML 19).

Precision-Recall \ AutoML \ метрика Gini

# MPP-challenge: моделирование прогноза качества модели

В своем докладе команда ВТБ подняла ряд важных вопросов касательно практического применения MPP-подхода в банковском моделировании:

1. Какие метрики ошибок первичных моделей лучше использовать как целевые переменные для MPP-моделей?

2. «Вручную» строить две модели слишком затратно по времени. Какие AutoML-подходы лучше работают для MPP-задач в кредитном скоринге и для других банковских задач?

3. Как работает MPP-мониторинг в периоды экономического стресса?

В этой статье мы попробуем ответить на поставленные вопросы, а также расскажем о нашем опыте разработки и внедрения в банке подхода MPP Renaissance Credit.

## Целевая переменная для MPP

Чтобы выбрать целевую переменную для MPP-модели, необходимо определить, по какой метрике качества будет проводиться мониторинг исходной модели.

Цель MPP-модели — строить прогнозы метрики качества исходной модели для невызревших кредитов.

Бóльшая часть задач банковского моделирования сводится к бинарной классификации, когда необходимо предсказать метки двух классов: попал кредит в дефолт или нет, откликнется клиент на кросс-сейл предложение или повесит трубку, «перекатится» клиент в худший бакет просрочки или внесет платеж, является транзакция нетипичной (мошеннической) или нет и т.д.

Для оценки качества бинарных классификаторов банки обычно используют коэффициент Gini (или родительскую метрику AUC-ROC), который рассчитывается как нормированная площадь между ROC-кривой исследуемой модели и ROC-кривой случайного классификатора (диагональю). Метрику Gini включают в мониторинги и наблюдают, как она меняется во времени.

Как уже было отмечено, в некоторых банковских моделях для расчета Gini требуется «вызревание» целевой переменной (например, просрочки). То есть на кредитах, выданных в последние несколько месяцев, расчет Gini становится невозможным из-за того, что просрочка по этим кредитам еще не «вызрела».

В этой ситуации можно использовать MPP-модель, цель которой — строить прогнозы метрики качества исходной модели для невызревших кредитов.

Однако для построения MPP-модели требуется сформировать выборку наблюдений с новыми метками классов (или числовыми целевыми значениями для задачи регрессии), которые отражают значения прогнозируемой метрики. И здесь возникает главная

проблема интегральных метрик (Gini, K-S, Average Precision и др.) — их нельзя привязать к отдельным наблюдениям выборки.

Решить эту проблему можно с помощью простого приема — подобрать пороговую метрику, которая наилучшим образом аппроксимирует поведение исходной интегральной метрики.

Например, для интегральной метрики Average Precision[1] наиболее близкой пороговой метрикой является F-мера (F1-score)[2], которая рассчитывается как среднее гармоническое пороговых метрик Precision и Recall (рис. 1а):

$$F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (1)$$

$$\text{Recall} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (3)$$

где TP (True Positive) — истинноположительные наблюдения;
FN (False Negative) — ложноотрицательные наблюдения;
FP (False Positive) — ложноположительные наблюдения.

В свою очередь пороговые метрики Precision и Recall являются координатами точек Precision-Recall кривой (при разных порогах), площадь под которой и есть интегральная метрика Average Precision. Таким образом, пороговая метрика *F*1-score и интегральная метрика Average Precision тесно связаны между собой (хотя эта связь зависит от выбранного порога).

Такой же «трюк» можно проделать и для интегральной метрики Gini, которая рассчитывается через площадь под ROC-кривой. Сама ROC-кривая — это совокупность точек с координатами пороговых метрик FPR и TPR. То есть мы можем взять среднее гармоническое этих двух координат и получить аналог *F*-меры для интегральной метрики Gini. Однако у полученной метрики будет один важный концептуальный недостаток. Компонента FPR имеет обратную связь с ранжирующей способностью модели, в то время как компонента TPR имеет прямую связь. Это значит, что при улучшении ранжирующей способности модели компонента FPR будет уменьшаться, а компонента TPR — увеличиваться, при этом их среднее гармоническое

[1] Другое название метрики Average Precision — AUC PR (площадь под PR-кривой).

[2] toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf.

# MPP-challenge: моделирование прогноза качества модели

будет меняться по параболической зависимости, то есть сначала расти, а потом убывать (или наоборот). Такая зависимость не соответствует математической интерпретации метрики Gini, которая растет с ростом ранжирующей способности модели.

Чтобы устранить этот недостаток, можно вместо компоненты FPR использовать обратный показатель TNR = Specificity = 1 – FPR, который имеет прямую зависимость с ранжирующей способностью модели. Тогда пороговая метрика для Gini будет иметь вид:

$$G\text{-score1} = 2 \times \frac{\text{TNR} \times \text{TPR}}{\text{TNR} + \text{TPR}}, \quad (4)$$

$$\text{TNR} = 1 - \text{FPR} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad (5)$$

$$\text{TPR} = \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (6)$$

где TN (True Negative) — истинноотрицательные наблюдения;
FP (False Positive) — ложноположительные наблюдения;
TP (True Positive) — истинноположительные наблюдения;
FN (False Negative) — ложноотрицательные наблюдения.

Сконструированная метрика *G*-score1 имеет прямую зависимость с Gini и отвечает требованиям нашей задачи[1].

Для дополнительных экспериментов мы сконструировали метрику *G*-score2, которая была получена из метрики *G*-score1 нормированием на корректирующий коэффициент Bad-Rate (долю *спрогнозированного* положительного класса):

$$G\text{-score2} = 2 \times \frac{\text{TNR} \times \text{TPR}}{(\text{TNR} + \text{TPR}) \times \text{BR}}. \quad (7)$$

Нормировка помогает сделать *G*-метрику более устойчивой, если Bad-Rate ведет себя нестабильно во времени (например, для моделей CRM).

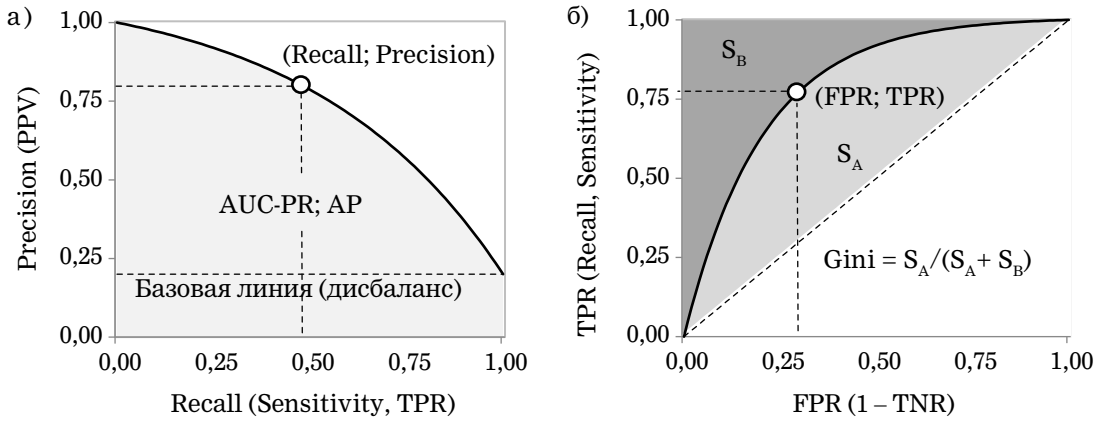
Для приведения метрик *G*-score1 и *G*-score2 к единому масштабу применяется нормировка значений к единичному диапазону, которая проводится на обучающей выборке.

Возвращаясь к постановке задачи MPP, теперь мы имеем пороговые метрики, которые полностью определяются четырьмя классами матрицы ошибок:

[1] Индекс в метрике *G*-score1 обозначает порядковый номер, в отличие от индекса в *F*1-score, где индекс обозначает степень в формуле усреднения компонент. Поэтому, чтобы не было путаницы, в обозначении *G*-score1 мы поставили индекс после слова score.

Рисунок 1

**Precision-Recall кривая и пороговые метрики Precision и Recall (а); ROC-кривая и пороговые метрики FPR и TPR (б)**




1) истинноположительные наблюдения (true positive — TP);
2) истинноотрицательные наблюдения (true negative — TN);
3) ложноположительные наблюдения (false positive — FP);
4) ложноотрицательные наблюдения (false negative — FN).

Таким образом, задача построения МРР-модели сводится к построению многоклассового классификатора, целевой переменной которого являются классы матрицы ошибок.

Важной задачей для формирования новой МРР-разметки является выбор порога (threshold) для расчета сконструированной метрики. Один из возможных методов выбора оптимального порога заключается в поиске максимального значения метрики, рассчитанной для исходной модели на различных порогах. Тогда точка, в которой метрика достигает максимального значения, принимается за оптимальный порог.

Также, помимо выбора целевой переменной МРР-модели и определения оптимального порога метрики, необходимо учитывать, что для разработки МРР-модели используется только тестовая выборка исходной модели, поскольку на обучающей выборке исходной модели нельзя получить *несмещенные* метки классов матрицы ошибок. Получается, что МРР-модель обучается на меньшем объеме данных, чем исходная модель, но при этом решает более сложную задачу. Если при обучении исходной модели применяется схема с k-fold

# MPP-challenge: моделирование прогноза качества модели

валидацией, то можно частично решить проблему нехватки данных, используя out-of-fold подвыборки для обучения MPP-модели.

В наших экспериментах мы проверяли качество работы метрик *G*-score1 и *G*-score2 на моделях скоринга (Application PD) и раннего взыскания (Collection model).

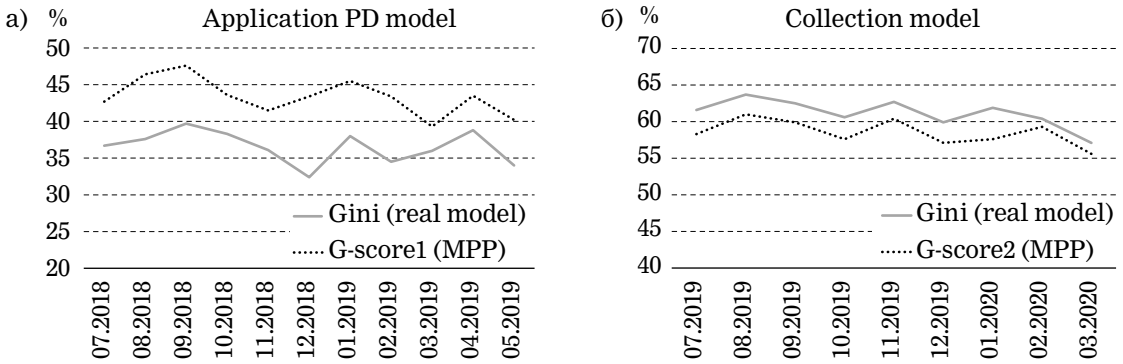
Для обучения MPP-моделей использовался градиентный бустинг в реализации LightGBM. Исходные модели обучались на логистической регрессии с использованием сплайнов.

На рис. 2 показаны результаты экспериментов на отложенной выборке out-off-time, которая не использовалась при обучении исходных моделей и MPP-моделей. На полученных результатах видно, что метрики *G*-score1 и *G*-score2, предсказанные MPP-моделями, хорошо аппроксимируют динамику изменения Gini для исходных моделей. Однако важно отметить, что качество аппроксимации MPP-модели зависит от качества исходной модели: чем ниже качество исходной модели, тем хуже MPP-аппроксимация (рис. 2а). И наоборот, чем выше качество исходной модели, тем лучше аппроксимация исходной метрики MPP-моделью (рис. 2б).

Задача построения MPP-модели сводится к построению многоклассового классификатора, целевой переменной которого являются классы матрицы ошибок.

Рисунок 2

**Прогноз MPP (метрика G-score1) для модели кредитного скоринга (а); прогноз MPP (метрика G-score2) для модели раннего взыскания (б)**




## AutoML для разработки MPP-моделей

Разработка моделей в КБ «Ренессанс Кредит» ведется в среде Python. Для обучения промышленных моделей мы используем единый автоматизированный Python-Pipeline, работающий в режиме End-to-End.

В промышленный Pipeline включены все ключевые этапы процесса разработки модели:

1) предобработка данных и подготовка выборок;

2) многоступенчатый отбор признаков;

3) настройка гиперпараметров и обучение модели (используемые алгоритмы: линейная регрессия, логистическая регрессия, сплайны, LightGBM);

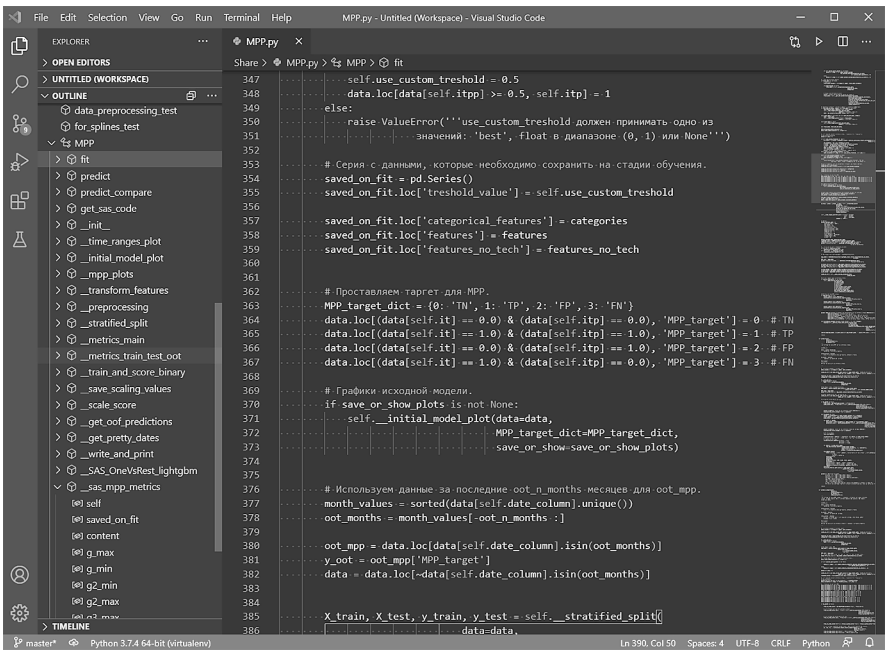
4) подготовка PMML-файлов для внедрения моделей в продакшн.

Единый Python-Pipeline позволяет автоматизировать весь процесс разработки моделей и масштабировать лучшие ML-решения, что позволяет снизить трудозатраты и модельные риски. Модульная структура пайплайна позволяет добавлять новые решения и удалять/заменять устаревшие.

Поскольку для построения MPP-моделей используются те же данные, что для исходных моделей, было принято решение встроить разработку MPP-моделей в наш промышленный пайплайн. Для этого был разработан MPP-модуль, который включает в себя 4 основных метода (рис. 3):

Рисунок 3

**MPP-модуль, встроенный в единый Python-Pipeline разработки промышленных моделей в КБ «Ренессанс Кредит»**



# MPP-challenge: моделирование прогноза качества модели

1) метод .fit() — обучает и сохраняет MPP-модель, выводит графики с метриками и статистиками;

2) метод .predict() — делает прогноз на основе обученной MPP-модели, выводит графики прогноза метрик;

3) метод .predict_compare() — выводит сравнение прогнозируемых и реальных метрик исходной модели;

4) метод .get_sas_code() — создает скрипт для внедрения расчета MPP-прогнозов в банковские мониторинги.

Разработанный MPP-модуль работает в режиме End-to-End и позволяет полностью автоматизировать все шаги обучения MPP-моделей и имплементации их в мониторинги.

## Работает ли MPP в период стрессов?

Проверку работы MPP-моделей в период стрессов мы проводили на выборках, захватывающих апрель-май 2020 г., когда в банковском сегменте наблюдался рост просроченной задолженности по розничным кредитам в связи с первой волной коронавируса в России.

MPP-модели способны угадывать изменения качества исходных моделей в периоды стрессов, однако качество MPP-прогнозов в периоды стрессов снижается.

Рабочая гипотеза заключалась в том, что в период стрессов MPP-модели могут плохо предсказывать падение качества исходных моделей по причине того, что MPP-модели обучались на данных без учета периодов экономического спада. Проверка MPP-прогнозов для моделей кредитного скоринга и раннего взыскания частично подтвердила это гипотезу (с поправкой на субъективность интерпретации).

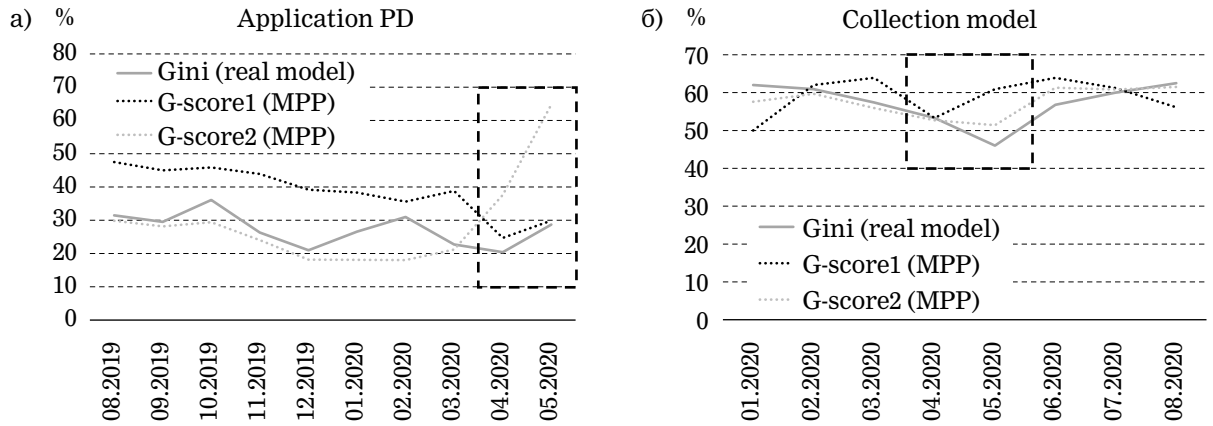
Для модели Application PD незначительное падение реального Gini наблюдалось в апреле 2020 г. (рис. 4а). Наиболее точно динамику Gini в этот период показала построенная на MPP-прогнозах метрика *G*-score1. Метрика *G*-score2, наоборот, не смогла предсказать падение качества исходной модели и показала рост в апреле 2020 г., а далее сильный рост в мае 2020 г.

Для модели Collection снижение Gini исходной модели наблюдалось в апреле-мае 2020 г., а в июне-июле качество модели вернулось на докризисный уровень. Наиболее точно эту динамику повторила метрика *G*-score2, а метрика *G*-score1 вела себя волатильно и не предсказала майский спад. При этом стоит отметить, что в период стресса (апрель-май) MPP-прогнозы имеют больше отклонение от реального Gini, чем в периоды до и после стресса.

Также при сравнении поведения MPP-прогнозов на разных моделях (рис. 4а–б) подтверждается гипотеза о том, что для исходных моделей с низким Gini прогнозы MPP-моделей также имеют низкое качество (нельзя построить сильную MPP-модель на слабых исходных признаках).

Рисунок 4

**Результаты прогноза MPP в кризисный период: MPP-прогнозы для модели кредитного скоринга (а); MPP-прогнозы для модели раннего взыскания (б)**




Можно заключить, что MPP-модели способны угадывать изменения качества исходных моделей в периоды стрессов, однако качество MPP-прогнозов в периоды стрессов снижается.

## Выводы и открытые вопросы

Подводя итоги, можно отметить, что MPP-модели способны предсказывать качество исходных моделей и могут быть использованы в мониторинге. При этом необходимо учитывать, что прогнозирующая способность MPP-моделей зависит от прогнозирующей способности исходных моделей (чем хуже качество исходной модели, тем ниже качество MPP-прогноза).

Применяя MPP-подход, необходимо также учитывать, что при разработке MPP-модели решается более сложная задача, чем при разработке исходной модели:

1) для обучения MPP-модели, как правило, используются более сложные ML-алгоритмы, такие как случайный лес или градиентный бустинг (в то время как для исходной модели из-за бизнес-ограничений может применяться более простая логистическая регрессия);

2) если исходная модель решает задачу бинарной классификации (прогнозирование двух классов), то для обучения MPP-модели приходится использовать четыре класса матрицы ошибок (т.е. увеличивается сложность классификации);

3) при построении MPP-модели требуется решить двойную задачу: предсказать прогнозы исходной модели и спрогнозировать исходную целевую переменную для расчета ошибки;

4) выборка, используемая для разработки MPP-модели, содержит значительно меньше признаков, чем выборка исходной модели, поскольку в MPP-выборку попадают только те признаки, которые вошли в исходную модель (число признаков может отличаться на порядок);

5) для разработки MPP-модели используется тестовая выборка исходной модели, то есть выборка для MPP-модели в несколько раз меньше, чем для исходной.

Перечисленные проблемы показывают, что, с одной стороны, разработка MPP-моделей требует более сложных подходов, с другой — для обучения MPP-моделей используется меньше данных, чем при обучении исходных моделей.

Несмотря на всю кажущуюся сложность, разработка MPP-моделей может быть полностью автоматизирована и включена в общий Pipeline разработки моделей. Написание Auto-ML модуля для автоматизации MPP-разработки занимает примерно 1–1,5 месяца работы ML-инженера.

Открытые вопросы, которые требуют дополнительного исследования:

1) точность MPP-прогнозов в период стрессов: улучшится ли качество прогнозов, если в MPP-выборку включать исторические стрессовые периоды?

2) определение наиболее оптимальных пороговых метрик для аппроксимации интегральных отраслевых метрик (Gini, K-S, Average Precision и др.).

Если говорить о дальнейшем развитии MPP-подхода в КБ «Ренессанс Кредит», то на данный момент мы начали строить MPP-модели для всех промышленных моделей банка, в планах расширенное тестирование MPP-подхода и исследование поведения построенных MPP-моделей.