Интепретируемая пороговая модель дохода клиента в банке

И. Емцев[I], Е.Р. Орлова[II]

[I] Базовая кафедра МФТИ при Федеральном исследовательском центре «Информатика и управление» Российской академии наук, г. Москва, Россия
[II] Федеральный исследовательский центр «Информатика и управление» Российской академии наук, г. Москва, Россия

Аннотация. В банковской индустрии точное и устойчивое прогнозирование доходов клиентов имеет ключевое значение для эффективного управления продуктами и рисковыми процессами. Существующие методы, включая классический градиентный бустинг[1] и вероятностные модели, нередко существенно недооценивают экстремальные значения, что приводит к систематическим ошибкам в хвостовой части распределения. В работе предлагается пороговая архитектура моделирования, основанная на раздельном прогнозировании тела и хвоста распределения с оптимальным выбором порога отсечения. Сравнение с традиционным бустингом, бустингом без «высоких доходов» и NGBoost демонстрирует существенное улучшение качества на крупных значениях дохода при сопоставимых результатах на основной массе данных. Методика обладает интерпретируемостью благодаря использованию SHAP и может служить эффективной альтернативой существующим моделям.

Ключевые слова: *прогнозирование дохода, банковская аналитика, ML, бустинг, экстремальные значения, GPD, пороговый метод, хвостовая аппроксимация, TIM.*
DOI: 10.14357/20790279260308 **EDN:** OLVOOD

Введение

Доход является одним из ключевых социально-экономических показателей, отражающих уровень жизни, потребительское поведение и финансовую устойчивость населения [1,2]. Для финансовых организаций точная оценка дохода критически важна, поскольку используется в кредитном скоринге, управлении рисками, ценообразовании продуктов и при выполнении регуляторных требований. При этом реальные доходы клиентов банков часто не полностью наблюдаемы, нерегулярны и подвержены значительным колебаниям, что усложняет задачу их моделирования.

Традиционные социально-демографические характеристики объясняют лишь ограниченную часть вариации доходов [3], а использование более детализированных данных существенно повышает точность предсказаний [4,5]. В то же время распределение доходов в банковских выборках обычно обладает выраженным тяжелым хвостом. Это приводит к тому, что единая модель, обученная на всем диапазоне доходов, даже при использовании современных *ML*-подходов, таких как градиентный бустинг, может недооценивать хвостовую часть распределения.

В актуарной литературе аналогичная проблема решается с помощью выделения тяжелых хвостов, что позволяет отдельно описывать экстремальные значения [6,7]. Эти подходы вдохновляют и предлагаемую архитектуру моделирования доходов (и на самом деле любой другой сущности с экстремальными значениями), где предполагается стратификация распределения на несколько взаимосвязанных компонентов. В отличие от классических *GLM*, традиционно используемых в страховании [8–10], мы предлагаем применить более современные методы машинного обучения, основанные на ***градиентном бустинге***, что позволяет эффективно работать с нелинейностями и сложными взаимодействиями признаков. Для интерпретации полученных моделей используется метод **SHAP**, обеспечивающий прозрачность и объяснимость предсказаний. Таким образом с помощью

[1] Градиентный бустинг – это метод машинного обучения. который создает мощные модели, комбинируя множество слабых предсказательных моделей, обычно деревьев решений.

использования идеи из [6,7] и адаптации ее под градиентный бустинг, можно получить значимый аплифт к качеству модели и при этом оставить ее интерпретируемой. Далее статья структурирована следующим образом.

Сначала описывается пороговая модель дохода (***Threshold Income Model***) и механизм формирования итогового предсказания. Затем рассматриваются модель для малых и средних доходов, модель для крупных доходов и модель вероятности превышения отсечки. После методологии приводится характеристика датасета, его особенности и метрики измерения качества. Далее описываются методы выбора отсечки на наших данных и результаты сравнения с текущими решениями. Также показывается вклад признаков с помощью SHAP. Статья завершается выводами, обсуждением ограничений и направлений дальнейших исследований.

1. Пороговая модель дохода

В модели дохода целевая переменная *Y* зачастую обладает сложной структурой: существенная масса наблюдений сосредоточена в районе малого и среднего дохода, в то время как хвост распределения представлен редкими, но большими значениями (например, высокодоходные клиенты, премиальные сегменты). Для получения устойчивой и интерпретируемой оценки распределения дохода целесообразно использовать пороговую модель, разделяющую тело и хвост доходов по фиксированной отсечке $u \in R_{>0}$. Далее в этом блоке будем считать, что отсечка $u \in R_{>0}$ является уже выбранной и зафиксированной. К ее выбору мы перейдем в секции с описанием датасета и применением модели.

Разделим распределение дохода *Y* по выбранной отсечке *u*. Для каждой тарифной ячейки[2] *r* пусть H_r и G_r – функции распределения доходов в теле и хвосте, соответственно, с параметрами θ_H и θ_G. Для обеспечения непрерывности распределения предположим:

$$H_r(u;\theta_H) > 0, \qquad G_r(u;\theta_G) = 0.$$

Пусть также $q_r(\theta_q)$ – вероятность превышения порогового значения *u* (вероятность попадания клиента в высокодоходный сегмент). Тогда функция распределения дохода в ячейке *r* с параметрами $\vec{\theta} = (\theta_H, \theta_G, \theta_q)$ имеет вид:

$$F_r\,(y;\theta) = \{0, y \le 0, (1 - q_r\,(\theta_q)) \cdot \\ \cdot (H_r\,(y;\theta_H))/(H_r\,(u;\theta_H)),$$

$$0 < y \le u, (1 - q_r\,(\theta_q)) + \\ + q_r\,(\theta_q) \cdot G_r\,(y;\theta_G), y > u.$$

Если h_r и g_r обозначают плотности H_r и G_r, соответственно, то плотность распределения дохода запишется как:

$$f_r\,(y;\theta) = \{0, y \le 0, (1 - q_r\,(\theta_q)) \cdot \\ \cdot (h_r\,(y;\theta_H))/(H_r\,(u;\theta_H)), 0 < y \le u, \\ q_r\,(\theta_q) \cdot g_r\,(y;\theta_G), y > u.$$

Определение 1. *Модель с функцией распределения и плотностью, указанными выше, называется Threshold Income Model (TIM).*

Таким образом, модель *TIM* является композицией трех подмоделей, разделенных пороговым значением *u*:

1. Модель для малого и среднего дохода – тело распределения (*H*);
2. Модель для высоких значений дохода – хвост распределения (*G*);
3. Вероятность превышения отсечки *u* (параметр *q*) – вероятность попадания клиента в хвост распределения.

Отсечка *u* считается заранее выбранной и фиксированной. Она не зависит от тарифной ячейки: редкость высокодоходных клиентов является глобальным свойством выборки. Если выбирать *u* отдельно по каждой клетке, модель будет склонна к переобучению и утрате интерпретируемости.

Таким образом математическое ожидание дохода в тарифной клетке *r* вычисляется как:

$$E(X \mid R = r) = \int_0^{+\infty} y \cdot f_r\left(y;\hat{\theta}\right) dy.$$

$$E(X \mid R = r) = \int_0^{+\infty} y \cdot f_r\left(y;\hat{\theta}\right) dy.$$

В статье [6] предлагают искать оценку параметров через метод максимального правдоподобия, мы же поступим по-другому и будем использовать предсказание бустингов, как итоговые значения промежуточных матожиданий: Введем $q_r = P(X > u \mid R = r)$ – вероятность превышения отсечки. Тогда можно записать разложение полного математического ожидания дохода по областям $X \le u$ и $X > u$:

$$E(X \mid R = r) \\ = P(X \le u \mid R = r)\, E(X \mid X \le u, R = r) \\ + P(X > u \mid R = r)\, E(X \mid X > u, R = r). \quad (1)$$

1.1. Модель для малого и среднего доходов ($X \le u$)

Так как распределение доходов в основной массе (до порога *u*) является сложно устроенным, нелинейным и является функцией множества при-

[2] Под тарифной ячейкой будем понимать агрегирующую категорию клиентов, например, комбинацию возраста клиента и его уровень риска.

знаков R, параметрическое описание здесь неэффективно и достаточно сложно. Поэтому условное математическое ожидание в области $X \le u$ аппроксимируется моделью градиентного бустинга. Обозначим ее прогноз через:

$$\hat{m}(r) := \hat{E}(X \mid X \le u, R = r) = Boost(r).$$

Бустинг напрямую приближает условное математическое ожидание, не предполагая фиксированной формы зависимости между признаками R и значением дохода в нижней части распределения.

1.2. Модель для крупных доходов ($X > u$)

Пусть есть клиенты $X_1, X_2, \ldots - i.i.d$ с функцией распределения F, которую мы не знаем. Запишем условную функцию распределения:

$$P(X > u + y \mid X > u) = \frac{1 - F(u + y)}{1 - F(u)},\ y > 0$$

Введем теорему для определения *Generalized Pareto Distribution (GPD)*

Теорема 1. *Пусть* $X_1, X_2, \ldots - i.i.d$ *с функцией распределения* F *и пусть для случайной величины* $M_n = \max\{X_1, \ldots, X_n\}$ *выполнено, что* $P(M_n \le z) \approx G(z) = \exp\left(-\left[1 + \xi \frac{z-\mu}{\sigma}\right]^{-\frac{1}{\xi}}\right)$, *для* $\mu > 0$ *и* $\sigma > 0$. *Тогда для достаточно больших и функция распределения случайной величины* $(X - u)$ *при* $X > u$:

$$P(X > u + y \mid X > u) = H(y) =$$

$$= 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}},\ \tilde{\sigma} = \sigma + \xi(u - \mu)$$

где $y \in \left\{y :\ y > 0\ \&\ \left(1 + \frac{\xi y}{\tilde{\sigma}}\right) > 0\right\}$

Определение 2. *H(y) является Generalized ParetoDistribution(GPD).Пусть* $x_{(i)} = \{x_i : x_i > u\}$, *тогда если* $y_{(i)} = x_{(i)} - u$, *тогда* $y \sim GPD(\sigma, \xi, u)$.

Значит

$$G_(\xi, \beta, u)\ (x) = \{\ 1 - (1 + \xi\ (x - u)/\beta\ )\text{^}(-1/\xi)\ ,\ \xi \ne 0\ \ 1 - \exp\left(-(x - u)/\beta\right)\ ,\ \xi = 0$$

$$g_{\xi,\beta,u}(x) = \frac{1}{\beta\left(1 + \xi\frac{x - u}{\beta}\right)^{-1-\frac{1}{\xi}}}$$

При этом параметры *GPD* уникально определяются параметрами $G(z)$ распределения. В частности ξ у обоих распределений будет одинаковым. При этом выбор параметра n не влияет на параметры *GPD*, если n достаточно большой (ξ не зависит от n, а μ и σ само компенсируются и не изменят $\tilde{\sigma}$).

Для моделирования хвоста будем использовать параметрическое распределение *GPD*: $g_{\hat{\xi},\hat{\beta},u}(y)$ с параметрами $\left(\hat{\xi}, \hat{\beta}, u\right)$, для которого условное ожидание имеет вид:

$$E(X \mid X > u, R = r) = \int_u^{r_I} y\ g_{\hat{\xi},\hat{\beta},u}(y)\,dy +$$

$$+ r_I\left(1 - G_{\xi,\beta,u}(r_I)\right),\quad r_I \ge u.$$

Последний член формально присутствует в выражении ожидаемого значения для неограниченных распределений, однако в банковских данных (особенно касается физических лиц) доход наблюдается только до какого-то максимального значения r_I, поэтому вклад этой части можно считать равной нулю. Для удобства далее используется известная аналитическая формула для матожидания *GPD*:

$$E(X \mid X > u, R = r) =$$

$$= \frac{\hat{\beta} - \hat{\xi}u + x}{\hat{\xi} - 1}\left(\frac{\hat{\xi}(x - u) + \hat{\beta}}{\hat{\beta}}\right)^{-1-\frac{1}{\hat{\xi}}}\Big|_u^{r_I}.$$

$$r_I \ge u.$$

Оценка параметра происходит с помощью выучивания распределения *GPD* на обучении с фиксированным $\mu = u$.

1.3. Модель для вероятности превышения отсечки (q_r)

Отдельным компонентом предлагаемой архитектуры является модель вероятности того, что доход клиента превышает пороговое значение u. Определим бинарную переменную

$$Y = 1_{\{X > u\}},$$

которая принимает значение 1, если доход клиента выше порога u, и 0 в противном случае. Тогда интересующая нас величина

$$q_r = P(X > u \mid R = r) = P(Y = 1 \mid R = r)$$

оценивается с помощью модели градиентного бустинга, решающей задачу бинарной классификации. Обозначим соответствующую модель через $Boost_{class}$. Тогда

$$\hat{q}_r := \hat{P}(X > u \mid R = r) = Boost_{class}(r),$$

где в качестве целевой переменной используется Y, а в качестве признаков – вектор тарифной клетки $R = r$. Модель $Boost_{class}$ обучается с использованием соответствующей функции потерь для классификации (например, логистической или бинарной кросс-энтропии) и возвращает для каждого r оценку вероятности превышения порогового дохода.

1.4. Итоговая формула для математического ожидания дохода

Подставляя модель бустинга для $E(X \mid X \le u, R = r)$, модель вероятности пре-

вышения отсечки $\hat{q}_r$, хвостовую модель для $E(X \mid X > u, R = r)$ в разложение (1), получаем окончательное выражение:

$$E_{\hat{\theta}}(X \mid R = r) = (1 - \hat{q}_r)\,\hat{m}(r)$$
$$+\hat{q}_r \frac{\hat{\beta} - \hat{\xi}u + x}{\hat{\xi} - 1}\left(\frac{\hat{\xi}(x-u) + \hat{\beta}}{\hat{\beta}}\right)^{-1-\frac{1}{\hat{\xi}}} |_u^{r_I}, r_I \geq u$$

Осталось только понять, как выбрать отсечку u, когда можно применять предположение, что $Y \sim GPD(\sigma, \xi, u)$.

2. Применение

Для повышения воспроизводимости и практической значимости исследования мы использовали реальный публичный датасет из соревнования *Kaggle Bank Credit Default (Loan Default Prediction)*, содержащий подробную социально-экономическую и транзакционную информацию о клиентах банка. Этот набор данных является одним из наиболее известных и репрезентативных примеров реальных банковских данных, доступных для академических и прикладных исследований.

Цель экспериментов – провести объективное сравнение четырех подходов (не включая линейные модели) к моделированию дохода клиента:

1. TIM – модель предложенная в статье;
2. Single Boosting – модель бустинга, обученная на всех наблюдениях (*baseline*);
3. Below Threshold – модель бустинга, обученная только на малых и средних доходах;
4. NGBoost – вероятностный метод градиентного бустинга, который обучает не только среднее значение целевой переменной, но и полное параметрическое распределение.

Для корректного и честного сравнения моделей фиксируются ключевые элементы эксперимента: гиперпараметры, набор признаков, разбиение на фолды (разобьем 1 раз на 5 фолдов) и все используемые значения *seed*. Также в качестве таргета будем брать не просто X, а $\log(X)$, так как это значительно улучшает процесс оптимизации функции потерь *RMSE* и делает обучение более устойчивым. Таким образом наш финальный ответ модели будет:

$$E_{\hat{\theta}}(X \mid R = r) = (1 - \hat{q}_r)\, e^{\hat{m}(r)}$$
$$+\hat{q}_r \exp\left(\frac{\hat{\beta} - \hat{\xi}u + x}{\hat{\xi} - 1}\left(\frac{\hat{\xi}(x-u) + \hat{\beta}}{\hat{\beta}}\right)^{-1-\frac{1}{\hat{\xi}}} |_u^{r_I}\right),$$
$$r_I \geq u.$$

2.1. Данные

Организаторы соревнования предоставляют исходные данные в виде двух выборок: обучающей (*train*) размером 307,511 наблюдений и тестовой (*test*) размером 48,744 наблюдения. Каждая запись соответствует одному клиенту и содержит десятки социально-демографических, поведенческих и продуктовых характеристик.

Поскольку не все признаки релевантны для задачи моделирования дохода и не все могут интерпретироваться в экономическом смысле, в работе используется отобранное подмножество признаков. В табл. 1 приведены наименования групп фичей (названия признаков сохраняются в исходном виде, принятом в соревновании).

Табл. 1

Категории используемых признаков

Категория признаков	Количество
Социально-демографические	3
Трудовая занятость	1
Региональные характеристики	3
Владение имуществом	3
Адресная согласованность	6
Контактная и верификационная информация	6
Поведенческие характеристики	1
Всего	**23**

Для упрощения эксперимента категориальные признаки были исключены, поскольку *NGBoost* не работает с ними напрямую, а необходимая предварительная обработка могла бы существенно исказить результаты моделирования. Также приведено распределение целевой переменной дохода (для наглядности выполнено *clip* до 1 млн, при этом значения выше данного порога составляют около 0,8% выборки) (рис. 1).

2.2. Метрики оценки качества

Для сравнения качества различных моделей дохода используются стандартные регрессионные метрики: корень из среднеквадратичной ошибки (*RMSE*), средняя абсолютная ошибка (*MAE*), а также показатель точности попадания предсказаний в заданный интервал (*hit rate, HRα*) [11]. Эти метрики вычисляются по следующим формулам:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,$$
$$HR_\alpha = \frac{1}{n}\sum_{i=1}^{n} I\{\hat{y}_i \in [(1-\alpha)y_i, (1+\alpha)y_i]\}.$$
$$HR_\alpha = \frac{1}{n}\sum_{i=1}^{n} I\{\hat{y}_i \in [(1-\alpha)y_i, (1+\alpha)y_i]\}.$$

Здесь y_i обозначает истинное значение дохо-

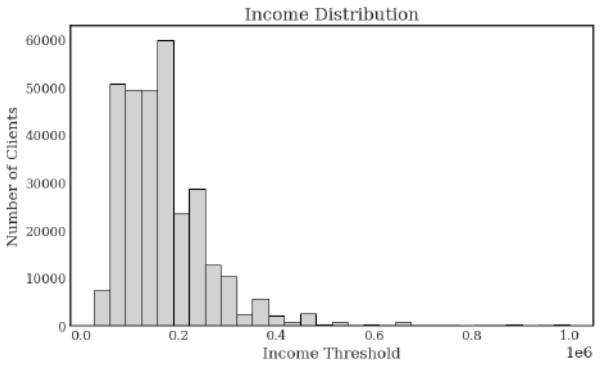


Рис. 1. Гистограмма распределения целевой переменной дохода AMT_INCOME_TOTAL

да, $\hat{y}_i$ – предсказанное значение, α задает допустимую относительную погрешность.

Особое внимание в данном исследовании уделяется метрикам **RMSE** и *HRα*. Метрика **RMSE** квадратично штрафует крупные ошибки и, тем самым, является наиболее чувствительной к отклонениям в области высоких доходов. Поскольку цель исследования заключается в корректном моделировании как основной массы распределения, так и его тяжелого хвоста, минимизация *RMSE* позволяет объективно оценить способность модели восстанавливать «крупные» значения дохода. Метрика *HRα* дополняет анализ, показывая по доходным бакетам, насколько точно модель попадает в соответствующие диапазоны и в каких областях ее ошибки превосходят ошибки альтернативных подходов. Для предлагаемой архитектуры (*TIM*) ожидается, что на малых и средних значениях доходов значения *HRα* будут сопоставимы с *baseline*, в то время как в области высоких доходов модель должна существенно превосходить все сравниваемые решения.

2.3. Выбор отсечки

Теперь перейдем к методам выбора отсечки на наших данных. Есть несколько способов выбора отсечки, предлагается их рассмотреть и с помощью их комбинации выбрать подходящую.

2.3.1. Первый способ выбора отсечки *u*

Если $Y \sim GPD(\sigma, \xi, u)$, то

$$EY = \{\sigma/(1-\xi)\ ,\ \xi < 1\ \infty\ ,\ \xi \geq 1$$

Пусть u_0 – минимальное значение отсечки, для которого начинает работать распределение *GPD*. Тогда распределение *GPD* работает и для $u > u_0$. Тогда матожидание можно представить в виде:

$$E(X - u \mid X > u) = \frac{\sigma_u}{1-\xi} = \frac{\sigma_{u_0} + \xi \cdot u}{1-\xi}.$$

Получается, что $E(X - u \mid X > u) \sim u$, т.е линейно от u.

Определение 3. $[E(X - u \mid X > u)$ – *mean of the excesses of the threshold u*

Определение 4:

$\left\{\left(u, \frac{1}{n_u}\sum_{i=1}^{(n_u)}\left(x_{(i)} - u\right)\right) : u < x_{max}\right\}$ n_u – *количество наблюдений выше отсечки – Mean Residual Life plot (рис. 2,а)*

Таким образом нам надо искать такие *u*, начиная с которых, график *Mean Residual Life Plot* ведет себя линейно, но надо помнить следующее и подбирать оптимальную *u*:

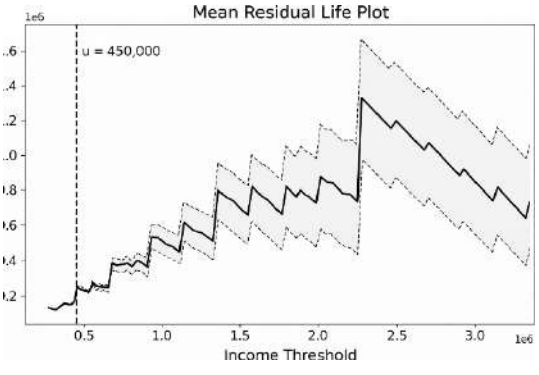
- Если *u* – очень маленькая, то получим смещенную оценку;
- Если *u* – очень большая, то получим очень широкую оценку,

Однако определение только по *Mean Residual Life Plot* является не самым приоритетным методом, поэтому перейдем ко второму способу.

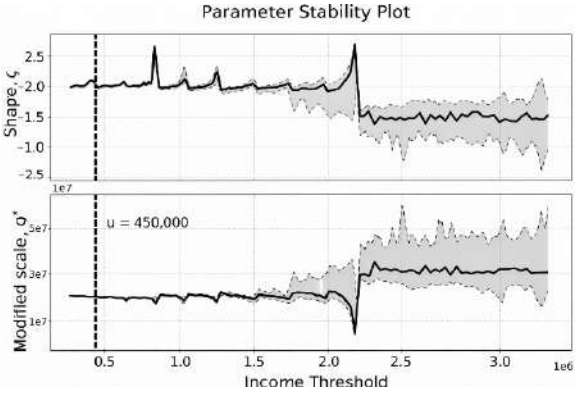
3.2.2. Второй способ выбора отсечки *u*

Поскольку $\sigma_u = \sigma_{u_0} + \xi \cdot (u - u_0)$, то если взять модифицированный scale параметр:

$$\sigma^* = \sigma_u - \xi \cdot u = \sigma_0 - \xi \cdot u_0 = constant = C \neq C(u).$$



а)



б)

Рис. 2. Mean residual life и Parameter stability на реальных данных

Табл. 2

Гиперпараметры для различных бустинговых моделей

	Название в Питоне	**CatBoost**	**CatBoost**	**NGBRegressor**
		Регрессия	Классификация	Регрессия
Model structure				
Number of trees	iterations	5000	5000	5000
Learning rate	learning_rate	0.05	0.05	0.05
Tree depth	depth	5	5	5
Loss / Distribution	loss_function	RMSE	Logloss	Normal
Metric	eval_metric	RMSE	AUC	LogScore
Regularization				
L2 regularization	l2_leaf_reg	4	4	–
Sampling parameters				
Row subsampling	subsample	0.9	0.9	0.9

Таким образом, когда работает распределение *GPD* мы имеем:

$$\{\sigma\hat{}* = C_1 \neq C_1(u)\,\xi =$$

$$= (\sigma_(u_0) - \sigma\hat{}*)/u_0 = C_2 \neq C_2(u).$$

На рис. 2,б мы можем увидеть стабильными σ^* и ξ. Основываясь на наши данные мы берем примерную оценку u_0=450,000. Так как мы работаем с $\log(X)$, то тогда берем отсечку $\log(u_0)$

3. Результаты

Для начала приведем пример гиперпараметров для всех моделей – они были выбраны плюс минус одинаковыми для всех архитектур для честной оценки аплифта каждой из моделей (табл. 2).

Табл. 3 демонстрирует качество предсказаний различных моделей в разрезе доходных интервалов, для которых модель попала в интервал ±5% от истинного значения (метрика HR_5). Такой анализ позволяет оценить поведение моделей как в низких и средних доходах, так и в зоне хвоста распределения, где расположены самые большие значения дохода.

Таким образом, в диапазоне доходов до 135*k* предложенная модель показывает качество лишь незначительно уступающее *baseline*. Разница составляет, как правило, не более 1–2 процентных пунктов. С учетом того, что основная часть выборки сосредоточена именно в этих интервалах, такая стабильность результатов указывает на то, что добавление пороговой структуры не приводит к заметной потере точности на массовом сегменте.

На интервалах выше 200*k* предложенная модель демонстрирует существенное превосходство над всеми альтернативными подходами. В этих бинах HR_5 возрастает от 50 до 100% относительно обычного бустинга, а по сравнению с моделью *Below-Threshold* превосходство становится многократным. Это подтверждает ключевое преимущество предлагаемого подхода: явное моделирование хвоста распределения позволяет корректно учитывать поведение редких высокодоходных наблюдений, что существенно повышает точность в правой части распределения.

Подход, основанный на обучении модели только на наблюдениях ниже порога, ожидаемо демонстрирует провал качества на высоких доходах. Зона правого хвоста практически не пред-

Табл. 3

Точность попадания в интервал ±5% по группам дохода на тесте

Интервал дохода	**TIM, %**	**Single Boosting, %**	**Below-Threshold,%**	**NGBoost,%**	**% выборки**
(27k–90k]	0.92	1.12	1.19	0.98	14.47
(90k–112k]	11.18	12.30	12.76	11.96	13.14
(112k–126k]	17.35	18.25	19.21	18.61	3.43
(126k–135k]	18.11	19.90	20.40	19.85	11.91
(135k–157k]	16.66	16.82	17.63	17.41	11.45
(157k–180k]	13.69	13.45	14.65	13.54	11.31
(180k–202k]	10.58	10.39	10.23	10.23	7.52
(202k–225k]	9.70	8.14	6.91	8.05	8.97
(225k–292k]	6.06	3.90	2.93	3.81	8.83
(292k–4.41M]	2.38	1.30	0.11	1.26	8.97

ставлена в данных, используемых для обучения этой модели, из-за чего она систематически недооценивает крупные значения. Падение HR_5 вплоть до 0.11% в самом высоком интервале четко указывает на неспособность модели корректно описывать поведение распределения доходов в области больших значений.

Несмотря на то, что *NGBoost* теоретически предназначен для аппроксимации полного распределения целевой переменной, на рассматриваемом датасете он демонстрирует качество ниже, чем обычный бустинг. Это наблюдается во всех бинах. Причинами могут быть высокая асимметрия и тяжелый хвост распределения доходов, которые плохо описываются нормальным распределением, используемым в *NGBoost*, а также нестабильность оптимизации параметров распределения по сравнению с минимизацией *RMSE* в градиентном бустинге.

Табл. 4

Сравнение моделей по RMSE и MAE с относительным улучшением (%) относительно худшей модели

Модель	RMSE	MAE
Test	baseline = 8.73×10^9	baseline = 56 329.7
TIM	(+5.65%)	(+0.62%)
Single Boosting	(+3.46%)	(+0.66%)
Below-Threshold	(0%)	(0%)
NGBoost	(+2.77%)	(+0.37%)
Avg K-Fold	baseline = 5.45×10^{10}	baseline = 54 688.9
TIM	(+0.68%)	(+0.55%)
Single Boosting	(+0.30%)	(+0.31%)
Below-Threshold	(0%)	(+1.00%)
NGBoost	(+0.21%)	(0%)

Помимо анализа метрик HR_5 по доходным бакетам, представленных ранее, следует отметить, что и по базовым качественным метрикам машинного обучения (**RMSE** и **MAE**) предлагаемая модель **TIM** демонстрирует практически наилучшие результаты среди всех сравниваемых подходов (табл. 4). Это означает, что **TIM** не только обеспечивает наиболее точное приближение распределения доходов в целом, но и остается стабильной как при наличии выбросов, так и при их отсутствии. Тем не менее, наибольшую практическую ценность для финансовых организаций представляет именно анализ HR_5 в разрезе доходных групп, поскольку он непосредственно отражает качество ранжирования клиентов по уровню дохода. Высокая точность попадания модели **TIM** в верхние бакеты позволяет выстраивать более надежные риск-стратегии и обеспечивает повышение качества нормативных расчетов, включая, например, правило показателя долговой нагрузки (**ПДН**).

Для демонстрации интерпретируемости применения бустингов для построения модели дохода [11] представлено распределение *SHAP*-значений, отражающих вклад отдельных признаков в итоговое предсказание дохода (рис. 4). Использование **SHAP** позволяет выявить ключевые факторы, определяющие вариацию модели, и тем самым подтверждает применимость бустинга в задачах, где требуется объяснимость результатов.

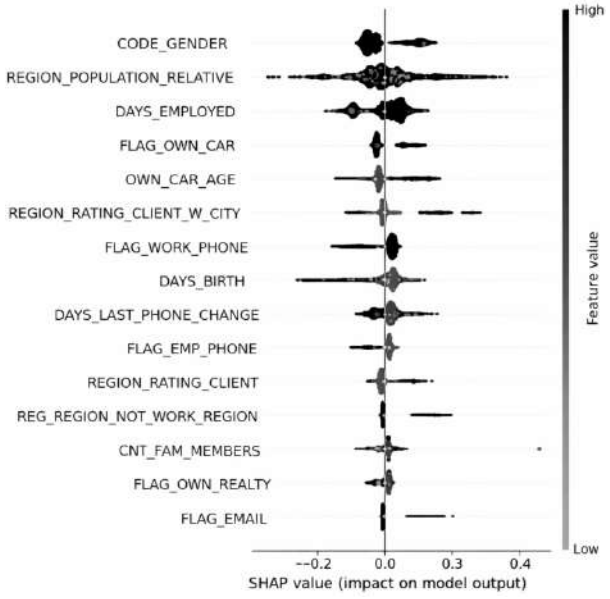


Рис. 4. Распределение SHAP-значений (top-15 признаков)

Краткая интерпретация части признаков:

1. FLAG_OWN_CAR – наличие автомобиля, как правило, коррелирует с более высоким уровнем дохода, что приводит к положительному вкладу данного признака в предсказание модели.
2. OWN_CAR_AGE – более новый автомобиль (меньший возраст) ассоциируется с более высоким доходом, в то время как значительный возраст автомобиля, напротив, снижает прогнозируемое значение.
3. DAYS_LAST_PHONE_CHANGE – недавняя смена номера телефона (значение ближе к нулю, учитывая отрицательную шкалу признака в выборке) может указывать на нестабильность поведения клиента, что приводит к снижению оценки его дохода.

Заключение

Представленная в работе пороговая архитектура моделирования доходов (**TIM**) демонстрирует принципиально иной подход к прогнозированию величин с выраженным тяжелым хвостом.

Разделение распределения на компоненты и специализированное моделирование его хвостовой части позволили существенно повысить точность предсказаний на высоких значениях дохода при сохранении стабильного качества на основной массе выборки. Такой результат подтверждает, что однородные модели, обученные на всем диапазоне данных, систематически проигрывают методам, учитывающим структурные особенности распределения.

Научная значимость предложенной методики заключается в создании общего фреймворка *Threshold Income Model (TIM)*, который не привязан к конкретному типу модели или предметной области. *TIM* может быть адаптирован к любой регрессионной задаче, в которой наблюдаются экстремальные значения: прогнозирование страховых выплат, моделирование потребления ресурсов, оценка операционных рисков, анализ больших транзакций и других количественных процессов с тяжелыми хвостами. В этом смысле методика расширяет классические идеи *Extrem Value Analysis (EVA)* и делает их совместимыми с современными *ML*-моделями.

Практическая значимость работы состоит в том, что *TIM* обеспечивает устойчивый прирост качества при сохранении интерпретируемости решений благодаря использованию *SHAP*. Это особенно важно в финансовом секторе, где требуется как высокая точность прогнозов, так и прозрачность моделей. Эксперименты показали, что при фиксированных гиперпараметрах, признаках и условиях обучения предложенный подход стабильно превосходит альтернативные решения.

Перспективы дальнейших исследований включают расширение методики на глубокие модели (*DL*-архитектуры), анализ динамических доходов во временных рядах, адаптацию *TIM* к многоуровневым распределениям и оценку его применимости в многомерных задачах, где требуется совместное моделирование хвостов нескольких переменных.

Литература

1. Ambrey C.L., Fleming C.M. The causal effect of income on life satisfaction and the implications for valuing non-market goods // Economics Letters. 2014. Vol. 123. No. 2. P. 131–134. DOI: 10.1016/j.econlet.2014.02.004
2. Burlacu M. The population's income, expenses and savings as descriptive aspects of the standard of living // Ovidius University Annals, Series Economic Sciences. 2016. Vol. 16. No. 2. P. 175–180. URL: https://stec.univ-ovidius.ro/html/anale/RO/2016/2016-II-full/s3/5.pdf (дата обращения: 10.02.2026)
3. Altonji J.G., Hynsjö D.M., Vidangos I. Individual earnings and family income: Dynamics and distribution // Review of Economic Dynamics. 2023. Vol. 49. P. 225–250. DOI: 10.1016/j.red.2022.08.005
4. Yu G.B. et al. Household income, satisfaction with standard of living, and subjective well-being. The moderating role of happiness materialism // Journal of Happiness Studies. 2020. Vol. 21. P. 2851–2872.
5. Kibekbaev A., Duman E. Benchmarking regression algorithms for income prediction modeling // Information Systems. 2016. Vol. 61. P. 40–52. DOI: 10.1016/j.is.2016.03.005
6. Laudagé C., Desmettre S., Wenzel J. Severity modeling of extreme insurance claims for tariffication // Insurance: Mathematics and Economics. 2019. Vol. 88. P. 77–92. DOI: 10.1016/j.insmatheco.2019.05.004
7. Coles S. An introduction to statistical modeling of extreme values. London: Springer, 2001. 209 p.
8. Ohlsson E., Johansson B. Non-life insurance pricing with generalized linear models. 2010. URL: https://www.irc.ac.ir/_douranportal/documents/tables/automobile/non-life%20insurance%20pricing.pdf (дата обращения: 10.02.2026)
9. Garrido J., Genest C., Schulz J. Generalized linear models for dependent frequency and severity of insurance claims // Insurance: Mathematics and Economics. 2016. Vol. 70. P. 205–215. DOI: 10.1016/j.insmatheco.2016.06.012
10. Schulz J. Generalized linear models for a dependent aggregate claims model: Master's thesis. Montréal: Concordia University, 2013. 87 p. URL: https://files01.core.ac.uk/download/pdf/211516813.pdf (дата обращения: 10.02.2026)
11. Salas P., Sáez P., Marchant V. An interpretable predictive model for bank customers' income using the eXtreme Gradient Boosting algorithm and the SHAP method: a case study of an Anonymous Chilean Bank // Research in Statistics. 2024. Vol. 2. No. 1. Art. 231229. DOI: 10.1080/27684520.2024.2312290

Емцев Илья. Базовая кафедра МФТИ при Федеральном исследовательском центре «Информатика и управление» Российской академии наук, г. Москва, Россия. Аспирант. Область научных интересов: цифровая трансформация. E-mail: emtsev.i@phystech.edu (Ответственный за переписку).

Орлова Елена Роальдовна. Федеральный исследовательский центр «Информатика и управление» Российской академии наук, г. Москва, Россия. Заведующая отделом. Доктор экономических наук, профессор. Область научных интересов: проблемы социально-экономического развития региона, инвестиционное развитие, цифровая трансформация. E-mail: orlova@isa.ru

Threshold Model for Estimating Bank Customers' Income

Ilya Yemtsev[I], E. R. Orlova[II]

[I] Moscow Institute of Physics and Technology' Basic Department at Federal State Institution "Federal Research Center "Computer Science and Management" of the Russian Academy of Sciences, Moscow, Russia

[II] Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

Abstract. In the banking industry, accurate and robust prediction of customer income is essential for effective product management and risk-related decision-making. Existing approaches, including classical gradient boosting and probabilistic models, often substantially underestimate extreme values, leading to systematic errors in the tail of the income distribution. This study proposes a threshold-based modeling architecture that separates the prediction of the body and the tail of the distribution with an optimally selected cutoff value. A comparison with traditional boosting, boosting without high-income observations, and NGBoost demonstrates a significant improvement in the accuracy of predicting large income values while maintaining competitive performance on the main part of the distribution. The proposed method remains interpretable through the use of SHAP values and can serve as an effective alternative to existing income prediction models.

Keywords: *income prediction, banking analytics, ML, boosting, extreme values, GPD, threshold method, tail approximation, TIM.*

DOI: 10.14357/20790279260308 **EDN:** OLVOOD

References

1. Ambrey C.L., Fleming C.M. The causal effect of income on life satisfaction and the implications for valuing non-market goods. Economics letters. 2014;123(2):131-134. https://doi.org/10.1016/j.econlet.2014.02.004
2. Burlacu M. The population's income, expenses and savings as descriptive aspects of the standard of living. Ovidius university annals, series economic sciences. 2016;16(2):175-180. Available from: https://stec.univ-ovidius.ro/html/anale/RO/2016/2016-II-full/s3/5.pdf [Accessed 10 February 2026]
3. Altonji J.G., Hynsjö D.M., Vidangos I. Individual earnings and family income: Dynamics and distribution. Review of economic dynamics. 2023;49:225-250. https://doi.org/10.1016/j.red.2022.08.005
4. Yu G.B., et al. Household income, satisfaction with standard of living, and subjective well-being. The moderating role of happiness materialism. Journal of happiness studies. 2020;21:2851-2872.
5. Kibekbaev A., Duman E. Benchmarking regression algorithms for income prediction modeling. Information systems. 2016;61:40-52. https://doi.org/10.1016/j.is.2016.03.005
6. Laudagé C., Desmettre S., Wenzel J. Severity modeling of extreme insurance claims for tariffication. Insurance: mathematics and economics. 2019;88:77-92. https://doi.org/10.1016/j.insmatheco.2019.05.004
7. Coles S. An introduction to statistical modeling of extreme values. London: Springer, 2001. 209 p.
8. Ohlsson E., Johansson B. Non-life insurance pricing with generalized linear models. 2010. Available from: https://www.irc.ac.ir/_douranportal/documents/tables/automobile/non-life%20insurance%20pricing.pdf [Accessed 10 February 2026]
9. Garrido J., Genest C., Schulz J. Generalized linear models for dependent frequency and severity of insurance claims. Insurance: mathematics and economics. 2016;70:205-215. https://doi.org/10.1016/j.insmatheco.2016.06.012
10. Schulz J. Generalized linear models for a dependent aggregate claims model [Master's thesis]. Montréal: Concordia University, 2013.

87 p. Available from: https://files01.core.ac.uk/download/pdf/211516813.pdf [Accessed 10 February 2026]
11. Salas P., Sáez P., Marchant V. An interpretable predictive model for bank customers' income using the eXtreme Gradient Boosting algorithm and the SHAP method: a case study of an Anonymous Chilean Bank. Research in statistics. 2024;2(1):231229. https://doi.org/10.1080/27684520.2024.2312290

Yemtsev Ilya. Postgraduate student at the MIPT basic department at FRC CSC RAS, Moscow, Russia. E-mail: emtsev.i@phystech.edu.

Orlova Eltna R. Doctor of Economics, Professor, Head of Department, Federal Research Center «Computer Sciences and Control» RAS, Moscow, Russia. E-mail: orlova@isa.ru