

АНАЛИЗ АНКЕТНЫХ ПАРАМЕТРОВ ЗАЕМЩИКОВ КАК ПРЕДИКТОРОВ ДЕФОЛТА ПОТРЕБИТЕЛЬСКИХ КРЕДИТОВ

Губин Е.И.¹, Франсис Н.Дж.², Старкова Ю.К.³

¹ Томский политехнический университет, к.ф.-м.н., доцент

² Томский политехнический университет, ИШИТР, Старший преподаватель, ОИТ ИШИТР

³ Томский политехнический университет, ИШИТР, студент гр. 8ПМ41, email: yks4@tpu.ru

Аннотация

Кредитный скоринг – один из ключевых инструментов для обеспечения стабильности финансовых учреждений и эффективного управления процессом потребительского кредитования, которое остается востребованным, но сопряжено с высокой вероятностью дефолта. В работе исследуется классификация заемщиков по анкетным данным для повышения точности оценки их кредитоспособности.

Ключевые слова: кредитный риск, градиентный бустинг, алгоритмы искусственного интеллекта.

Введение

Целью данной работы является сравнение моделей классификации, предназначенных для оценки влияния отдельных признаков на вероятность дефолта заемщика по потребительскому кредиту. В исследовании используются современные методы машинного обучения CatBoost и LightGBM. Методология анализа опирается на принципы, изложенные в научных трудах по кредитному скорингу и управлению рисками, таких как «Advances in Financial Machine Learning» (M. Lopez de Prado) [1].

Описание алгоритма

Для решения поставленной задачи были применены алгоритмы машинного обучения. Исходным набором данных служил датасет, содержащий информацию о 3000 заемщиках и 27 признаках, включая возраст, профессию, доходы, количество членов семьи, детей, количество завершенных кредитов, размер запрашиваемой суммы кредита, регион проживания, а также продолжительность трудовой деятельности и времени проживания по месту регистрации. В процессе анализа данных были выявлены выбросы, пропущенные значения, а также признаки, имеющие высокую корреляцию между собой и не оказывающие значимого влияния на модель. Пропущенные значения были заменены на медианы, выбросы обработаны с использованием метода IQR, а высоко коррелирующие признаки были исключены. Также была применена кросс-валидация для повышения надежности модели. Строки с «плохими» заемщиками были помечены как 1, а строки с «хорошими» - как 0.

Перед нами стояла задача классификации заемщиков, для этого были выбраны алгоритмы градиентного бустинга CatBoost из пакета catboost и LightGBM Classifier из пакета lightgbm. Важность признаков для обеих моделей оценивалась с помощью функции feature_importances [3].

CatBoost использует градиентный бустинг для построения модели, где каждое новое дерево исправляет ошибки предыдущего, а важность признаков рассчитывается на основе их вклада в улучшение модели через разбиения дерева. CatBoost автоматически обрабатывает категориальные признаки, что значительно повышает его эффективность. LightGBM, применяя гистограммное разбиение признаков, ускоряет обучение, снижая вычислительные расходы. Важность признаков в LightGBM вычисляется на основе их частоты использования в разбиениях и их вклада в уменьшение ошибки модели.

В Таблице 1 и на Рис.1 показаны признаки из анкетных данных, которые наиболее сильно влияют на дефолт по версии модели CatBoost.

Таблица 1. Важность признаков по версии CatBoost

Признак	Важность (%)
CASH	9.677
TMJOB1	9.511
TMADD	8.476
INCOME	6.8534
PERS_H	6.5367

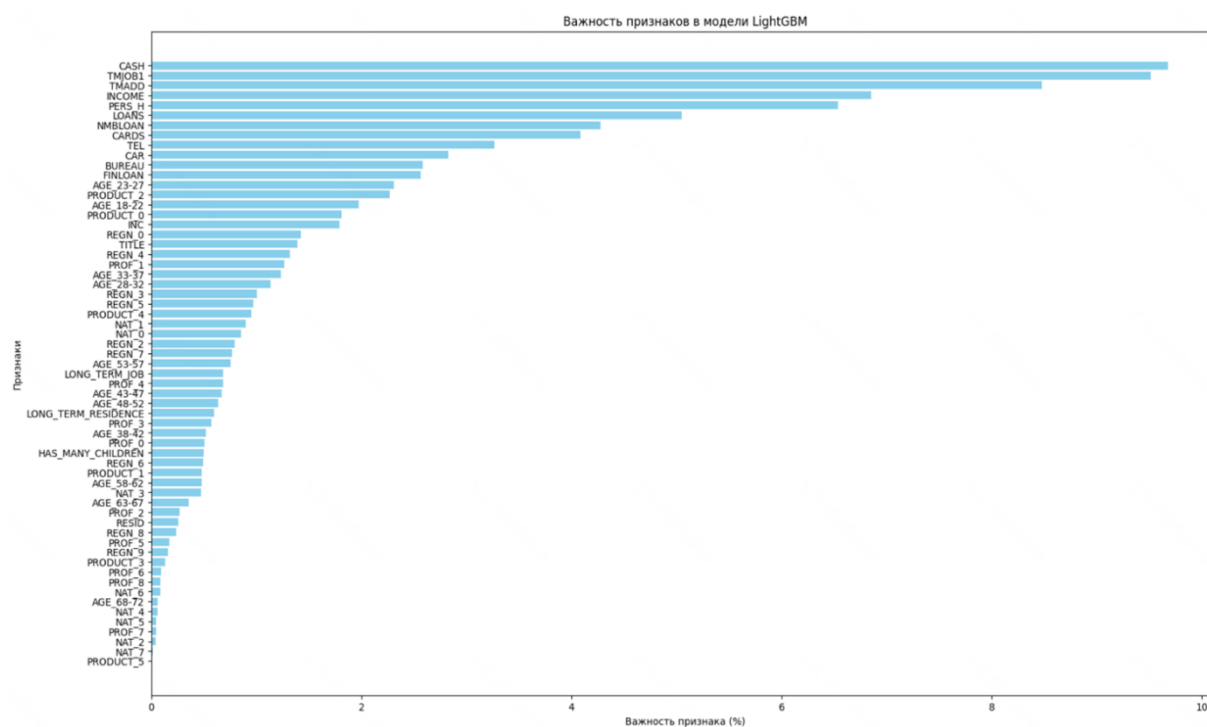


Рис.1. Важность признаков для предсказания дефолта по версии CatBoost

LightGBM выявила аналогичные признаки, которые приведены в таблице 2 и на рис. 2. Однако на первом месте TMJOB1 вместо CASH.

Таблица 2. Важность признаков по версии LightGBM

Признак	Важность (%)
TMJOB1	14.135
CASH	13.4
TMADD	12.198
INCOME	11.243
PERS_H	5.12

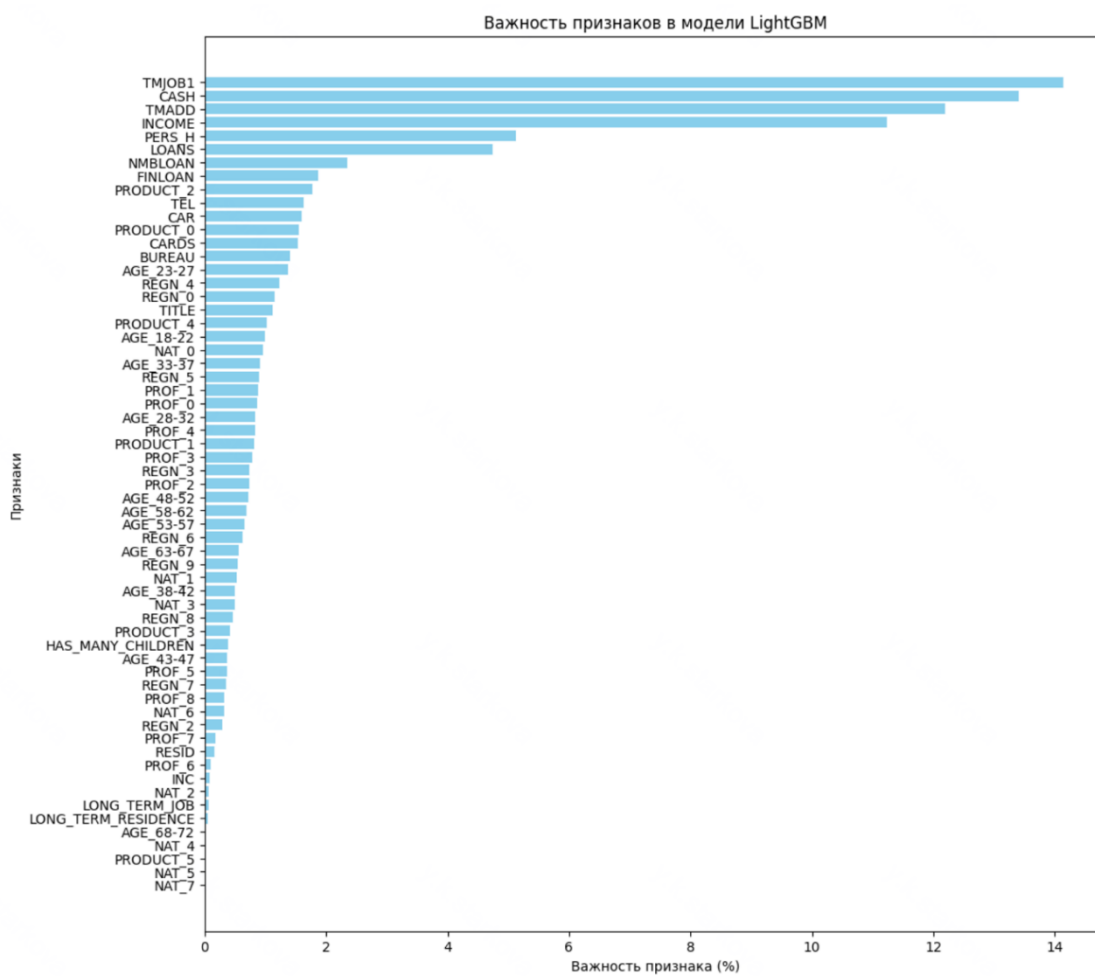


Рис.2. Важность признаков для предсказания дефолта по версии LightGBM

Показатели метрик классификатора [2] приведены в таблице 3. Точность модели для LightGBM составила 0.77 %, в случае CatBoost – 73 %.

Таблица 3. Показатели метрик классификатора

Metric	LightGBM (1)	LightGBM (0)	CatBoost (1)	CatBoost (0)
Precision	0.77	0.73	0.73	0.75
Recall	0.71	0.79	0.66	0.80
F1-score	0.74	0.76	0.695	0.77
Support	2000	1000	2000	1000

Результаты анализа показали, что LightGBM имеет более высокую точность (77 %), что свидетельствует о более стабильной работе модели в общем контексте. В отличие от LightGBM, модель **CatBoost продемонстрировала более высокое значение recall для дефолтных заемщиков**, что указывает на **лучшую способность этой модели выявлять потенциальных неплательщиков**. Однако за счет этого **общая точность модели оказалась ниже**, поскольку увеличилось число ложных срабатываний, то есть случаев, когда добросовестные заемщики ошибочно классифицировались как дефолтные.

LightGBM использует гистограммный метод разбиения, который ускоряет обучение и позволяет находить более точные разбиения на малых данных. Это может давать ему преимущество в условиях ограниченного объема выборки.

CatBoost строит симметричные деревья, что способствует более стабильному обучению. Однако этот подход снижает гибкость модели, из-за чего алгоритм хуже справляется с общей точностью, но лучше определяет заемщиков с высоким риском дефолта.

Заключение

Если основной целью является максимизация точности модели (например, минимизация ложных срабатываний для добросовестных заемщиков), то предпочтительным будет использование LightGBM. Однако если задача заключается в минимизации пропуска дефолтных заемщиков, то CatBoost может быть более эффективным инструментом. В будущем возможны исследования по улучшению моделей, включая размножение данных с помощью генетических алгоритмов, создание новых признаков, а также использование нейронных сетей и эксперименты с методами обработки выбросов и заполнения пропусков.

Список использованных источников

1. López de Prado M. *Advances in Financial Machine Learning*. Wiley. – 2018.
2. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. – М : 2016. – 2017. – 393 с.
3. Документация Scikit-learn. [Электронный ресурс]. – URL: scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (дата обращения: 21.01.2025)