

# Отчет по проекту: Применение ансамблей градиентных бустингов в задаче кредитного скоринга

Аманкулов Айдар Бакытбекович  
amankulov.a@phystech.edu

10 Декабрь, 2024

## 1 Литературный обзор

### 1.1 Введение в кредитный скоринг

Кредитный скоринг является crucial инструментом в современной банковской системе, позволяющим оценивать кредитоспособность заемщиков. Согласно исследованию Hand и Henley [1], скоринговые модели позволяют существенно снизить риски невозврата кредитов и автоматизировать процесс принятия решений.

### 1.2 Традиционные методы скоринга

Исторически для кредитного скоринга применялись:

- Логистическая регрессия
- Дискриминантный анализ
- Деревья решений

Thomas et al. [2] отмечают, что логистическая регрессия долгое время оставалась стандартом де-факто благодаря своей интерпретируемости и надежности.

### 1.3 Развитие ансамблевых методов

Brown и Mues [3] провели сравнительный анализ различных методов машинного обучения в задачах кредитного скоринга, показав преимущества ансамблевых методов над одиночными моделями.

#### 1.3.1 Градиентный бустинг

Chen и Guestrin [4] представили XGBoost – реализацию градиентного бустинга, которая стала прорывом в области машинного обучения. Ke et al. [5] разработали LightGBM, оптимизированный для работы с большими данными.

#### 1.3.2 Преимущества градиентного бустинга в скоринге

Исследования Lessmann et al. [6] показали, что градиентный бустинг обладает рядом преимуществ:

- Высокая предсказательная способность
- Устойчивость к выбросам
- Способность работать с пропущенными данными
- Автоматическое выделение важных признаков

## 1.4 Современные подходы к оптимизации

Akiba et al. [7] представили Optuna – фреймворк для автоматической оптимизации гиперпараметров, который значительно упростил настройку сложных моделей. Zhang и Zhou [8] исследовали различные стратегии ансамблирования, включая: Стекинг (stacking), Блендинг (blending), Взвешенное голосование.

## 1.5 Проблемы и вызовы

Современные исследования выделяют следующие актуальные проблемы:

Интерпретируемость моделей (Molnar [9]), Дисбаланс классов (He и Garcia [10]), Временная стабильность моделей (Hand [11]).

## 2 Библиография

- [1] Hand, D.J., Henley, W.E. (1997) "Statistical Classification Methods in Consumer Credit Scoring: A Review"
- [2] Thomas, L.C., et al. (2002) "Credit Scoring and its Applications"
- [3] Brown, I., Mues, C. (2012) "An experimental comparison of classification algorithms for imbalanced credit scoring data sets"
- [4] Chen, T., Guestrin, C. (2016) "XGBoost: A Scalable Tree Boosting System"
- [5] Ke, G., et al. (2017) "LightGBM: A Highly Efficient Gradient Boosting Decision Tree"
- [6] Lessmann, S., et al. (2015) "Benchmarking state-of-the-art classification algorithms for credit scoring"
- [7] Akiba, T., et al. (2019) "Optuna: A Next-generation Hyperparameter Optimization Framework"
- [8] Zhang, C., Zhou, Y. (2016) "Ensemble Machine Learning: Methods and Applications"
- [9] Molnar, C. (2019) "Interpretable Machine Learning"
- [10] He, H., Garcia, E.A. (2009) "Learning from Imbalanced Data"
- [11] Hand, D.J. (2006) "Classifier Technology and the Illusion of Progress"

Напиши используя Latex код, но немного покороче

## 3 Актуальность проекта

В современном финансовом секторе задача оценки кредитоспособности заемщиков является одной из ключевых проблем управления рисками. Актуальность исследования методов кредитного скоринга обусловлена несколькими важными факторами:

### 1. Экономическая значимость:

- Рост объемов кредитования требует автоматизации процессов оценки заемщиков
- Увеличение числа невозвратных кредитов создает необходимость в более точных методах оценки рисков
- Экономические кризисы повышают важность качественного скоринга

### 2. Технологические аспекты:

- Развитие методов машинного обучения открывает новые возможности для повышения качества моделей
- Градиентный бустинг показывает высокую эффективность в задачах классификации
- Появление новых источников данных требует современных методов их обработки

### 3. Практическая значимость:

- Снижение рисков финансовых организаций
- Ускорение процесса принятия решений по кредитным заявкам
- Повышение качества кредитного портфеля

В контексте этих факторов, исследование применения ансамблей градиентных бустингов в задаче кредитного скоринга представляется особенно актуальным, так как позволяет значительно повысить точность предсказаний и автоматизировать процесс принятия решений при оценке кредитоспособности заемщиков.

## 4 Постановка задачи

Целью данного исследования является сравнительный анализ эффективности различных алгоритмов градиентного бустинга в задаче кредитного скоринга, а также исследование возможностей повышения качества предсказаний с помощью методов стекинга и блендинга.

### 4.1 Формальная постановка

Пусть имеется набор данных  $D = \{(x_i, y_i)\}_{i=1}^n$ , где:

- $x_i \in \mathbb{R}^d$  - вектор признаков  $i$ -го заемщика
- $y_i \in \{0, 1\}$  - метка класса (0 - надежный заемщик, 1 - дефолт)
- $n$  - количество наблюдений
- $d$  - размерность пространства признаков

### 4.2 Основные задачи

1. Предобработка данных:

$$X_{\text{processed}} = f_{\text{preprocess}}(X_{\text{raw}})$$

2. Оптимизация параметров моделей:

$$\theta^* = \arg \max_{\theta} \text{ROC-AUC}(M_{\theta}(X_{\text{val}}), y_{\text{val}})$$

где  $M_{\theta}$  - модель с параметрами  $\theta$

3. Построение ансамбля моделей:

$$F_{\text{ensemble}}(x) = \sum_{k=1}^K w_k M_k(x)$$

где  $M_k$  - базовые модели,  $w_k$  - их веса

### 4.3 Критерии качества

Основные метрики для оценки качества моделей:

- ROC-AUC:  $\text{AUC} = \int_0^1 \text{TPR}(t) \text{FPR}'(t) dt$

## 5 Теоретическая часть

### 5.1 Градиентный бустинг

Градиентный бустинг представляет собой ансамблевый метод машинного обучения, основанный на последовательном построении композиции алгоритмов:

$$F_M(x) = \sum_{m=1}^M \gamma_m h_m(x)$$

где:

- $F_M(x)$  - итоговая модель
- $h_m(x)$  - базовые алгоритмы (чаще всего решающие деревья)
- $\gamma_m$  - коэффициенты
- $M$  - количество итераций

## 5.2 Реализации градиентного бустинга

### 5.2.1 XGBoost

Оптимизирует следующую целевую функцию:

$$\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

где  $\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$  - регуляризационный член.

### 5.2.2 LightGBM

Использует технику GOSS (Gradient-based One-Side Sampling):

$$\tilde{g}_i = \begin{cases} g_i, & \text{для важных примеров} \\ \frac{a}{b} g_i, & \text{для остальных} \end{cases}$$

### 5.2.3 CatBoost

Применяет упорядоченный бустинг:

$$\text{Target}_i = \frac{\sum_{j=1}^{i-1} [y_j = 1]}{\sum_{j=1}^{i-1} 1}$$

## 5.3 Методы ансамблирования

### 5.3.1 Стекинг

$$f_{\text{final}}(x) = g(f_1(x), f_2(x), \dots, f_K(x))$$

где  $g$  - метамодель,  $f_k$  - базовые модели.

### 5.3.2 Блендинг

$$f_{\text{blend}}(x) = \sum_{k=1}^K w_k f_k(x)$$





где  $w_k$  - веса моделей.

## 6 Практическая часть

Весь этап практической части исследования проходил в ноутбуке формата `ipynb`. Ссылку на этот ноутбук прикладываю тут:

<https://colab.research.google.com/drive/1RlyJ5gqbwrddqA71-tcaXRJF2ByzErjJ7?usp=sharing>

## 7 Скриншоты с соревнования с kaggle.com

 <b>xgboost_optuned(1).csv</b> Complete (after deadline) · 3h ago	<b>0.72094</b>
 <b>lgbm_optuned(2).csv</b> Complete (after deadline) · 2h ago	<b>0.72277</b>
 <b>blending_xgb_rf_lgbm(1).csv</b> Complete (after deadline) · 13h ago	<b>0.72277</b>
 <b>stacking_xgb_without_categ.csv</b> Complete (after deadline) · 1h ago	<b>0.72571</b>