Hackerearth ML Challenge 1
Bank Fears Loanliness

Евгений Патеха

Конкурс на Hackerearth
ML Challenge 1. Bank Fears Loanliness

- Задача – предсказание вероятности дефолта по выданным займам
- Объем данных:
  - Train – **532 428** записей
  - Test – **354 951** записей
  - Метрика – **AUC**

public lb – 50%, private lb – 50%

1.	**Evgeny Patekha**	**0,98650**
2.	Divanshu Garg	0,98329
3.	Yuvraj Rathore	0,97945
4.	Bhupinder	0,97927
5.	Mario Filho	0,97708
6.	Nikita Trifonov	0,97707
7.	vopani	0,97657
8.	João Paulo Vasques Camargo da Silva	0,97654
9.	Alex Chernobrovov	0,97633
10.	ziron	0,97627

44 признака

- Параметры займа
- Финансовое состояние клиента
- Данные об обслуживании

Variable	Description
loan_status	status of loan amount
member_id	unique ID assigned to each member
batch_enrolled	batch numbers allotted to members
loan_amnt	loan amount ($) applied by the member
term	term of loan (in months)
int_rate	interest rate (%) on loan
sub_grade	grade assigned by the bank
home_ownership	status of home ownership
annual_inc	annual income ($) reported by the member
verification_status	status of income verified by the bank
purpose	purpose of loan
tot_cur_bal	total current balance of all accounts
revol_bal	total credit revolving balance
revol_util	% of credit a member is using
total_rec_int	interest received till date
total_rec_late_fee	Late fee received till date
recoveries	post charge off gross recovery
total_acc	total number of credit lines available
open_acc	number of open credit line

Код клиента – случайный носитель информации

Валидация

- Данные поделены на батчи, сильно различающиеся по средней вероятности дефолта (от 1% до 90+%)
- В test и train разные батчи
- Кросс-валидация с обычным разбиением на stratified фолды без учета батчей может привести к переобучению
- Данные на фолды разбивались по принципу все строки одного батча в один фолд плюс равномерное распределение записей и положительных статусов дефолта между фолдами

Анализ прироста результата

Новые фичи	CV (5 folds)*	SD (5 folds)
base	0,974014	0,00104
add total_rec_int_to_funded & total_rec_int_to_funded_by_sgrade	0,982746	0,00059
add total_rec_int_to_funded_to_term	0,983223	0,00052
add last_week_pay_by_batch & last_week_pay_by_id300k	0,984111	0,00093
add int_r_subgrade	0,984351	0,00074
add batch_count	0,984378	0,00078
add annual_inc_prec	0,984454	0,00086

* Результаты кросс-валидации в lightGBM с learning_rate .05
Результаты на leaderboard в xgboost с learning_rate .01 выше примерно на .002

Подготовка данных и новые признаки

Подготовка данных

- Перевод текстовых признаков в числа (статусы, грейды, стаж работы и тд), заполнение NA и пустых признаков -1

Отбор признаков

- Добавление новых признаков по одному и оценка прироста через кросс-валидацию

Лучшие новые признаки

```
dt[, total_rec_int_to_funded:= total_rec_int / (funded_amnt*int_rate)]

dt[, total_rec_int_to_funded_by_sgrade:= (total_rec_int_to_funded-
mean(total_rec_int_to_funded)) / sd(total_rec_int_to_funded), by=.(sub_grade)]
```

Вычислительные ресурсы. ПО

- Ноутбук - 2 ядра, 12 GB
- Google Cloud – 8 ядер 16 GB (финальные сабмишены с низким eta)
- R пакеты data.table, xgboost, lightGBM

Спасибо за внимание!