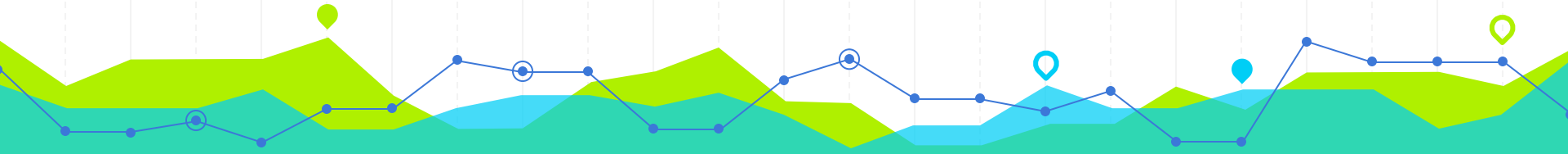
Прогнозирование попадания вагонов в текущий ремонт

Кирилл Захаров
4 курс, СПбГЭУ
Прикладная математика и информатика в экономике и управлении

Содержание работы

1. Анализ данных
2. Построение моделей
3. Сравнение результатов



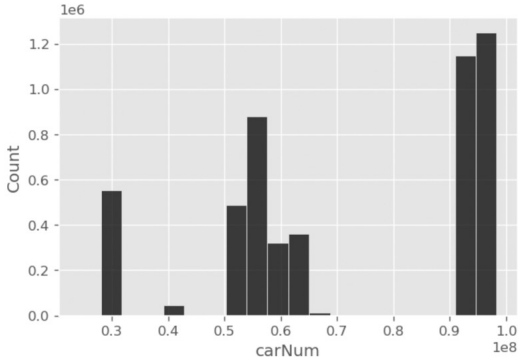
Анализ данных

1

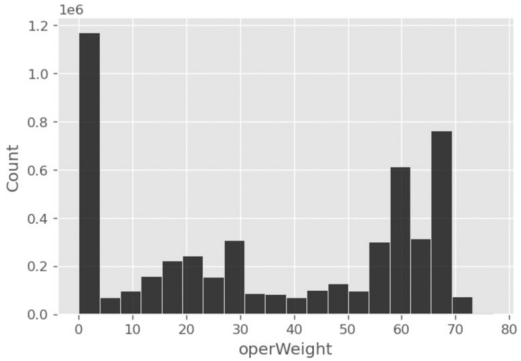
Дислокации

- Номер вагона
- Дата дислокации (год, месяц, день, время)
- Код операции
- Код станции
- Код станции назначения
- Код груза
- Вес груза

Записи о вагонах



Распределение весов



Основные коды операций: ОТПР, ПРИБ, РМНТ

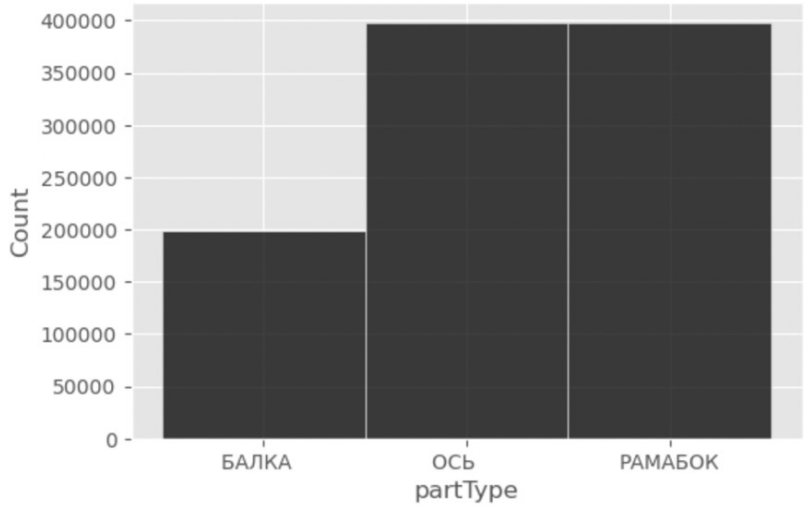
Ремонты

- Номер вагона
- Тип ремонта (текущий, деповской, капитальный)
- Дата начала ремонта
- Дата окончания ремонта
- Код депо
- Код железной дороги
- Название станции

Детали для ремонта вагонов

- Номер вагона
- Дата поступления вагона
- Деталь, которая заменяется
- Депо установки детали
- Дата установки детали
- Код депо установки детали
- Год производства детали
- Характеристики для «Осей»

Количество записей о деталях



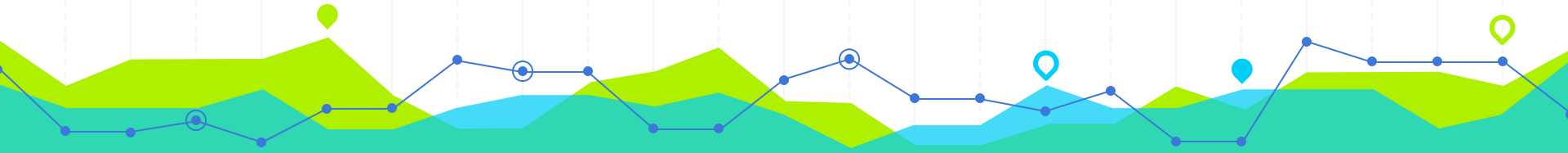
Гребни

- Номер вагона
- Дата замера
- Толщина гребня каждой пары колес

axle1_rf	axle1_lf	axle2_rf	axle2_lf	axle3_rf	axle3_lf	axle4_rf	axle4_lf
27.8	25.6	26.9	27.1	27.3	25.7	27.1	25.8
26.7	29.7	28.8	28.5	26.0	26.6	26.4	29.5
28.2	27.3	27.1	29.8	23.9	27.1	29.3	26.9

Формирование единой таблицы данных

- Номер вагона
- Дата дислокации
- Код дислокации
- Код станции назначения
- Код груза
- Вес груза
- Глубина обода
- Ширина гребня (после последнего ремонта)
- Количество деталей по типу («Рамабок», «Ось», «Балка»)
- Ширина гребня (в текущий момент времени)



Построение моделей

2

Модели

1. Logistic Regression
2. Random Forest (400 деревьев)
3. Gradient Boosting (100 деревьев)
4. AdaBoost based on Random Forest
5. Ensemble 1 (LR, RF, GB)
6. Cat Boost
7. Ensemble 2 (LR, ABRF, CB)

Метрики

Полнота (Recall)

$$R = \frac{TP}{TP + FN}$$

Точность (Precision)

$$P = \frac{TP}{TP + FP}$$

F1-score

$$F1 = 2\frac{P * R}{P + R}$$

F-beta (beta=5)

$$F_\beta = (1 + \beta^2)\frac{P * R}{\beta^2 P + R}$$

macro-average

$$mavg = \frac{R_0 + R_1}{2}$$

weighted-average

$$wavg = \frac{R_0 \#0 + R_1 \#1}{\#0 + \#1}$$

Logistic Regression

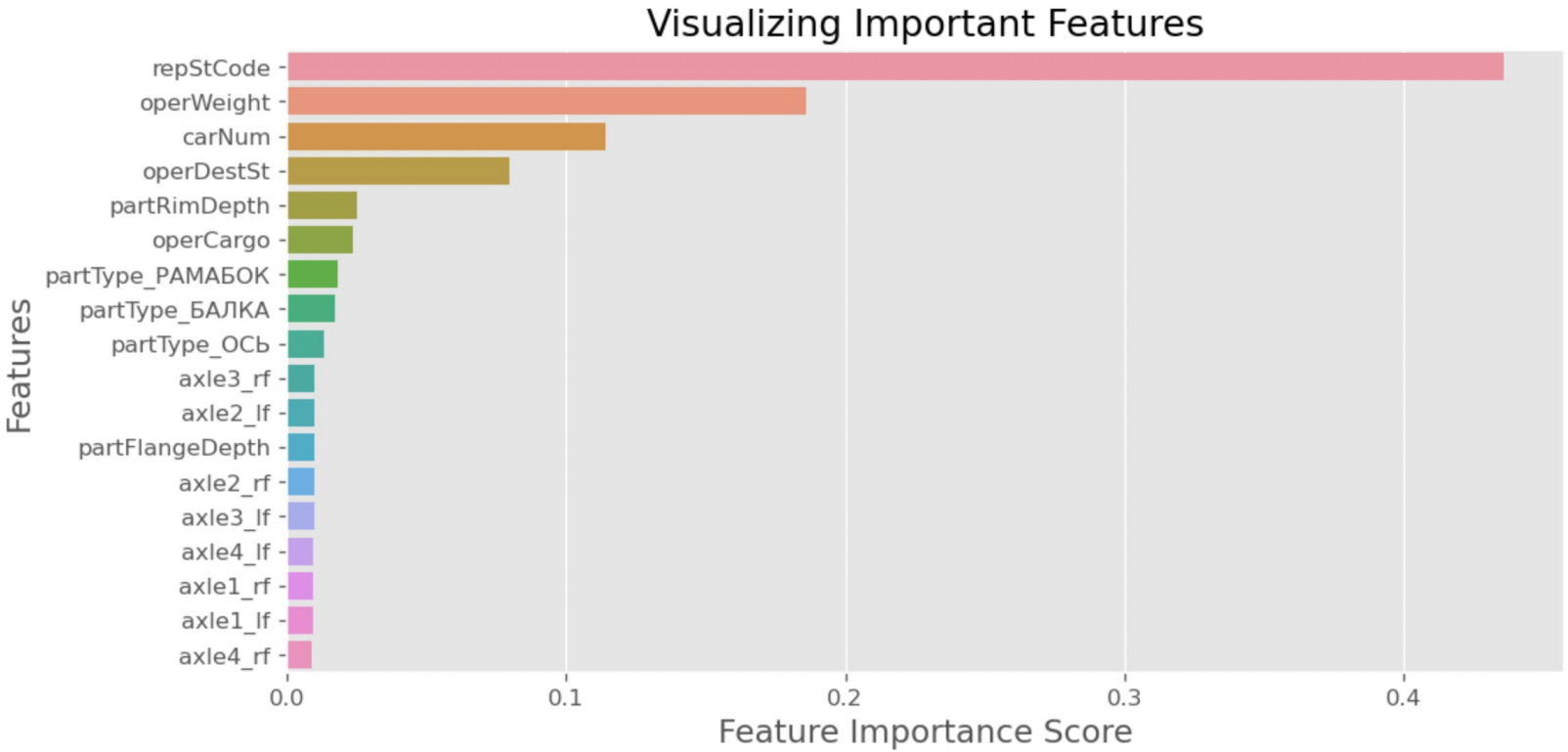
	precision	recall	f1-score	support
0	1.00	0.83	0.91	739262
1	0.01	0.85	0.02	1210
accuracy			0.83	740472
macro avg	0.50	0.84	0.46	740472
weighted avg	1.00	0.83	0.90	740472

Random Forest

Деревьев: 400
Глубина: 10

	precision	recall	f1-score	support
0	1.00	0.97	0.98	739262
1	0.04	0.84	0.08	1210
accuracy			0.97	740472
macro avg	0.52	0.90	0.53	740472
weighted avg	1.00	0.97	0.98	740472

Random Forest



Gradient Boosting

Деревьев: 100
Глубина: 3
lr: 0.1

	precision	recall	f1-score	support
0	1.00	1.00	1.00	739262
1	0.53	0.38	0.44	1210
accuracy			1.00	740472
macro avg	0.76	0.69	0.72	740472
weighted avg	1.00	1.00	1.00	740472

Ada Boost based on Random Forest

Деревьев: 50
Глубина: 10
n_esimators: 10

	precision	recall	f1-score	support
0	1.00	0.99	1.00	739262
1	0.12	0.71	0.21	1210
accuracy			0.99	740472
macro avg	0.56	0.85	0.60	740472
weighted avg	1.00	0.99	0.99	740472

Ensemble 1 (LR+RF+GB)

	precision	recall	f1-score	support
0	1.00	0.99	0.99	739262
1	0.08	0.71	0.15	1210
accuracy			0.99	740472
macro avg	0.54	0.85	0.57	740472
weighted avg	1.00	0.99	0.99	740472

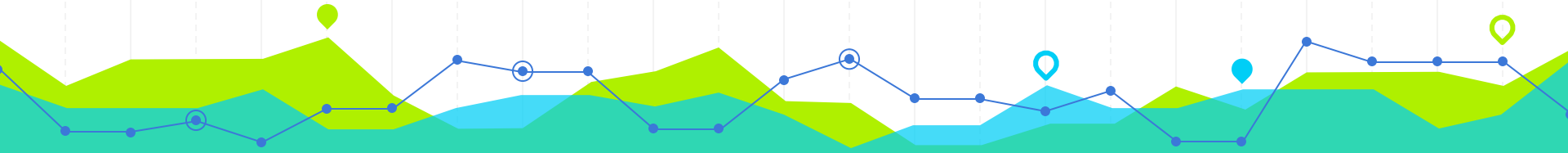
Cat Boost

iterations: 100
Глубина: 15
lr: 0.1

	precision	recall	f1-score	support
0	1.00	1.00	1.00	739262
1	1.00	0.38	0.55	1210
accuracy			1.00	740472
macro avg	1.00	0.69	0.77	740472
weighted avg	1.00	1.00	1.00	740472

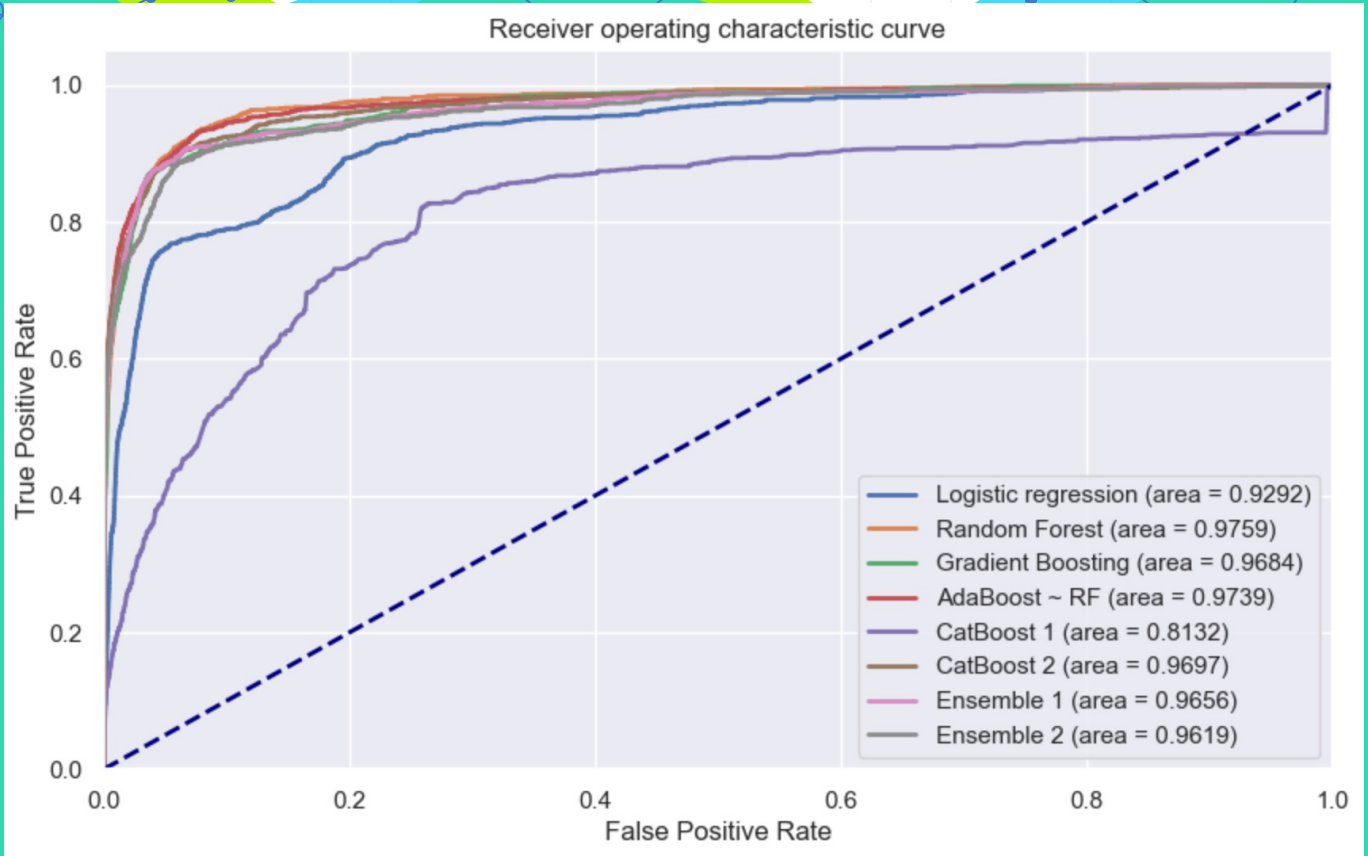
Ensemble 2 (LR+ABRF+CB)

	precision	recall	f1-score	support
0	1.00	1.00	1.00	739262
1	0.21	0.60	0.31	1210
accuracy			1.00	740472
macro avg	0.61	0.80	0.66	740472
weighted avg	1.00	1.00	1.00	740472

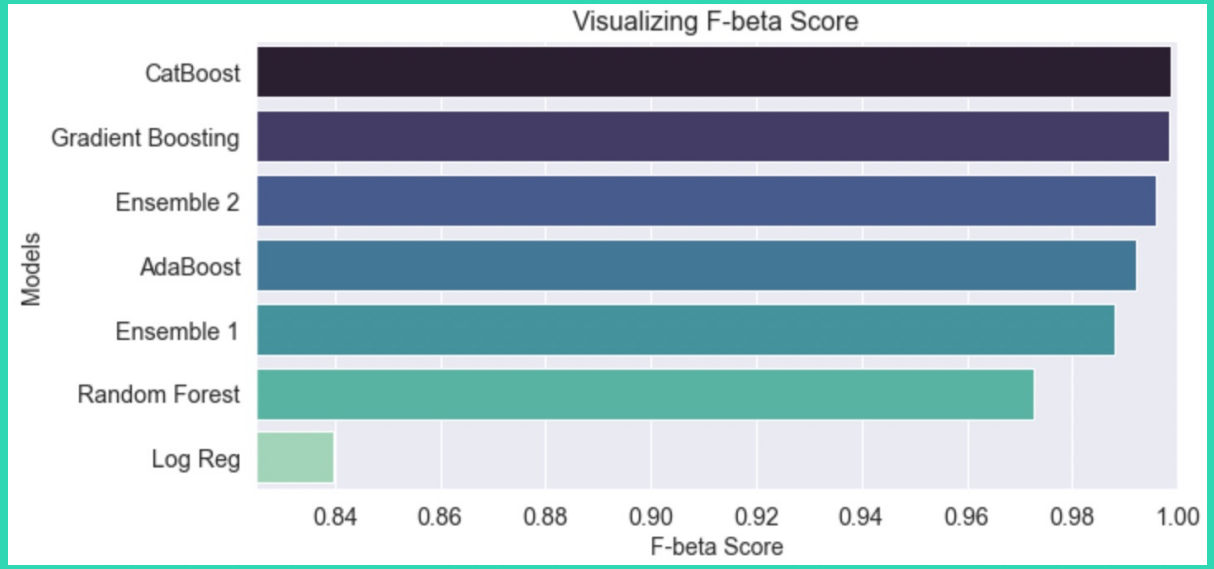


Сравнение результатов 3

ROC curves и AUC scores



F-beta score



CatBoost	0.998868
Gradient Boosting	0.998370
Ensemble 2	0.995921
AdaBoost	0.992203
Ensemble 1	0.987908
Random Forest	0.972629
Log Reg	0.839676

Сравнение по всем метрикам

Модели	Precision	Recall	F1	F-beta	macro avg (R)	weighted avg (R)	auc
LR	0.01	0.85	0.02	0.8396	0.84	0.83	0.929
RF	0.04	0.84	0.08	0.9726	0.90	0.97	0.975
GB	0.53	0.38	0.44	0.9883	0.69	1	0.968
ADRF	0.12	0.71	0.21	0.9922	0.85	0.99	0.973
E1	0.08	0.71	0.15	0.9879	0.86	0.99	0.965
CB	1	0.38	0.55	0.9988	0.69	1	0.969
E2	0.21	0.60	0.31	0.9959	0.8	1	0.962

Прогнозирование вероятностей

carNum	Station	Probability	Actual
28061984	27200.0	0.320125	0
	76610.0	0.324074	0
	27373.0	0.192203	0
	26700.0	0.236876	0
	27230.0	0.202894	0
...	...	...	...
98077332	83283.0	0.242398	0
	23060.0	0.444396	0
	25823.0	0.258210	0
	79040.0	0.318437	0
	25442.0	0.248559	0

carNum	Station	Probability	Actual
28061984	27200.0	0.320125	0
	76610.0	0.324074	0
	27373.0	0.192203	0
	26700.0	0.236876	0
	27230.0	0.202894	0
	76060.0	0.286568	0
	64000.0	0.817927	1
	26720.0	0.181641	0
	27144.0	0.194322	0
	27140.0	0.178632	0
	26600.0	0.219815	0
	26770.0	0.235186	0
	27230.0	0.331121	0
	24580.0	0.217399	0

Спасибо за внимание!