Задание 1. Предсказание значений целевой переменной для тестовой выборки на основе исторических данных по совершениям транзакций клиентов.

Для начала изучили данные.

Общее количество уникальных клиентов, по которым предоставлена история транзакций = 12000.

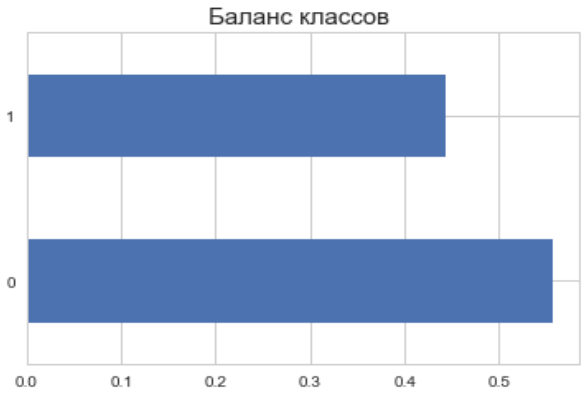

Рисунок 1 – Баланс классов.

Баланс классов в обучающей выборке = 0.556222/ 0.443778. Выборка оказалась сбалансированной.

В качестве исходных данных было представлено поле **datetime**, включающее в себя номер дня, а также время совершения транзакций. Для того, чтобы вынести информацию из данного показателя, а именно, год, месяц и день совершения транзакции, была просмотрена динамика количества платежных транзакций (транзакции, в которых поле SUM имеет отрицательный знак) по каждому номеру дня в разрезе кода транзакции 5992(code = 5992), то есть динамика количества дневных платежных транзакций по всем клиентам для кода «Флористика».

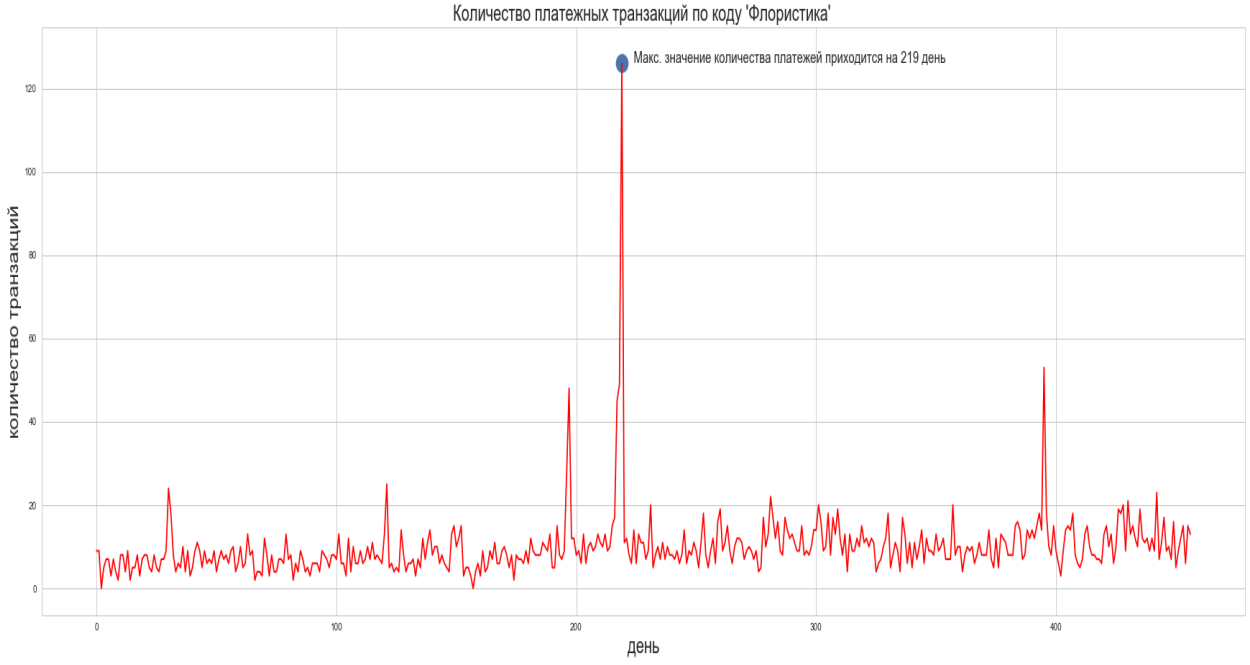

Рисунок 2 – Динамика количества дневных транзакций по коду “Флористика”.

Из рисунка 2 видно, что наибольшее число платежных транзакций по коду «Флористика» приходится на 219 день. Исходя из этого, можно предположить, что это 8 марта. Пик, наблюдающийся рядом –это 197 день, скорее приходящийся на ближайший праздник- 14 февраля. Делаем вывод, что 219 день не только является 8 мартом, но и то, что год, приходящийся на этот праздник- не високосный, так как разница между 8 мартом и 14 февраля равна 20 дней.

Преобразовав номер дня совершения транзакции в дату в формате (YYYY:MM:DD) построим дневные, месячные и годовые агрегаты для исходных показателей.

Построение агрегатов

Создаем поле-индикатор, которое будет отвечать за период времени совершенной транзакции- рабочий день или ночь (рабочий день в период с 09:00 – 21:00, ночь – в оставшийся период времени)

Также создаем поле-индикатор расходной или доходной транзакции абонентов(из поля SUM)

Из исходных полей type и code формируем новые показатели со статистикой COUNT - встречаемости определенного кода и типа транзакции для каждого клиента в разрезе всей истории.

Годовые, месячные и дневные агрегаты

Для каждого клиента генерируются следующие показатели:

- Из исходных полей type и code формируем новые показатели со статистикой COUNT - встречаемости определенного кода и типа транзакции для каждого клиента в разрезе года, месяца, дня.
- Суммарная доходная транзакция по каждому клиенту
- Суммарная платежная транзакция по каждому клиенту
- Остаток на счете за расчетный период(день, месяц, год)
- MIN, MAX, MEAN, STD величины доходной транзакции
- MIN, MAX, MEAN, STD величины платежной транзакции
- Количество платежных, доходных транзакций
- Количество дневных, ночных транзакций

Таким образом, сгенерировали 4171 полей из исходных показателей для 12000 клиентов.

Препроцессинг:

- Замена пропусков на 0;
- Удаление полей, содержащих одно уникальное значение
- Корреляционный анализ(порог коэфициента корреляции Пирсона = 0,985)

Выбор алгоритма

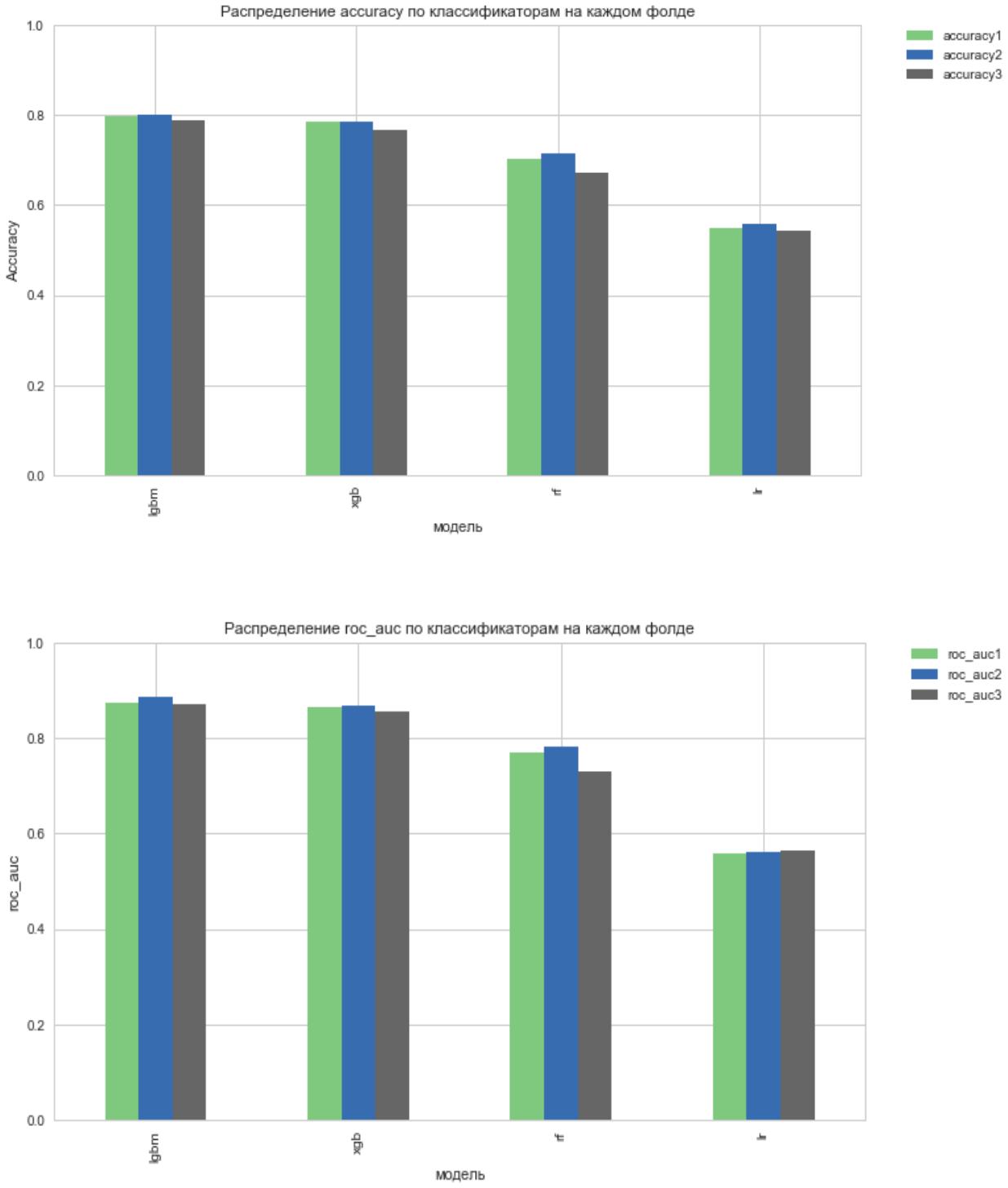

Рисунок 3 – Качество моделей

На графиках выше показаны метрики accuracy и roc_auc для каждого, лучшего в своем роде алгоритма на стратифицированной кросс-валидации.

Из графиков видно, что наилучшим и стабильным алгоритмом является **LightGBM** со следующими параметрами:

```
LGBMClassifier(boosting_type='gbdt', class_weight=None,
    colsample_bytree=1.0,
    learning_rate=0.01, max_depth=-1, metric='roc_auc',
    min_child_samples=45, min_child_weight=0.001, min_split_gain=0.0,
    n_estimators=1000, n_jobs=-1, num_leaves=60, objective=None,
    random_state=None, reg_alpha=0.0, reg_lambda=0.0, silent=True,
    subsample=0.7, subsample_for_bin=2000, subsample_freq=1)
```

В данном случае оптимизационной метрикой выбрал roc-auc.

Качество модели:

Качество нашей модели неплохое и стабильное(без переобучения).

```
cross_validation roc-auc score on 3 folds =
[ 0.87321219  0.88509115  0.87143936]
roc_auc_mean on 3 folds equal to =  0.8765809001357868


cross_validation accuracy score on 3 folds =
[ 0.79773409,  0.80033333,  0.78692898]
accuracy_mean on 3 folds equal to =  0.79499879943196528
```

Важность показателей(top(30)):

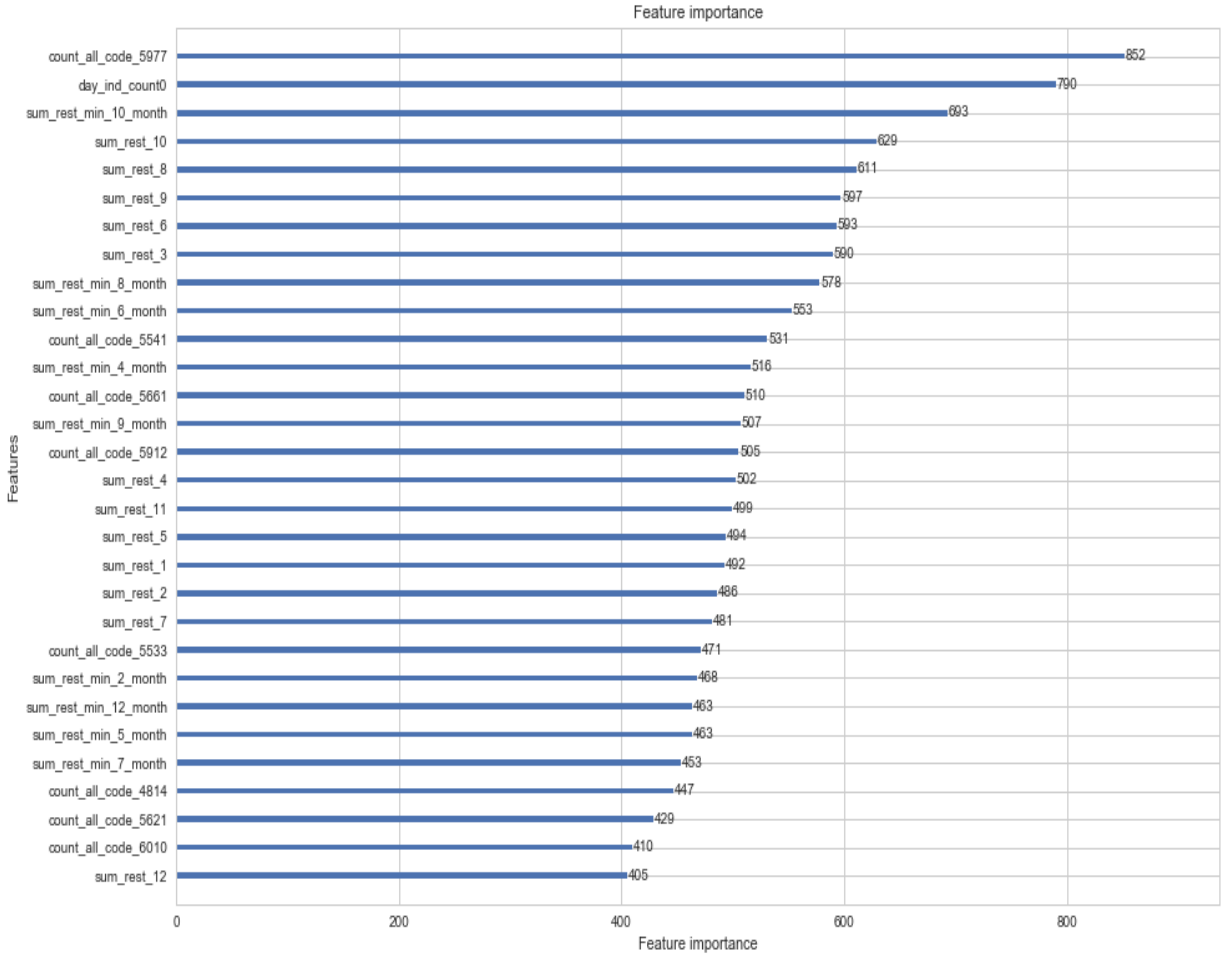

Рисунок 4 – важность переменных

Расшифровка наиболее значимых показателей

Название показателя	Расшифровка
count_all_code_5977	Количество транзакций по коду «Магазины косметики» за весь период истории
day_ind_count0	Количество ночных транзакций за весь период
sum_rest_min_10_month	Минимальный остаток на счете за 10 месяц(величина минимальной доходной транзакции за вычетом соответствующей величины платежной транзакции за 10 месяц)
sum_rest_10	Отсаток на счете за 10 месяц
sum_rest_8	Отсаток на счете за 8 месяц
sum_rest_9	Отсаток на счете за 9 месяц
sum_rest_6	Отсаток на счете за 6 месяц
sum_rest_3	Отсаток на счете за 3 месяц
sum_rest_min_8_month	Минимальный остаток на счете за 8 месяц(величина минимальной доходной транзакции за вычетом соответствующей величины платежной транзакции за 8 месяц)
sum_rest_min_6_month	
count_all_code_5541	Количество транзакций по коду «Станции техобслуживания» за весь период истории
…	…

Выбор отсечки:

Ниже показан рисунок, показывающий, как отсечка по скорбаллу влияет на качество классификации через метрику accuracy

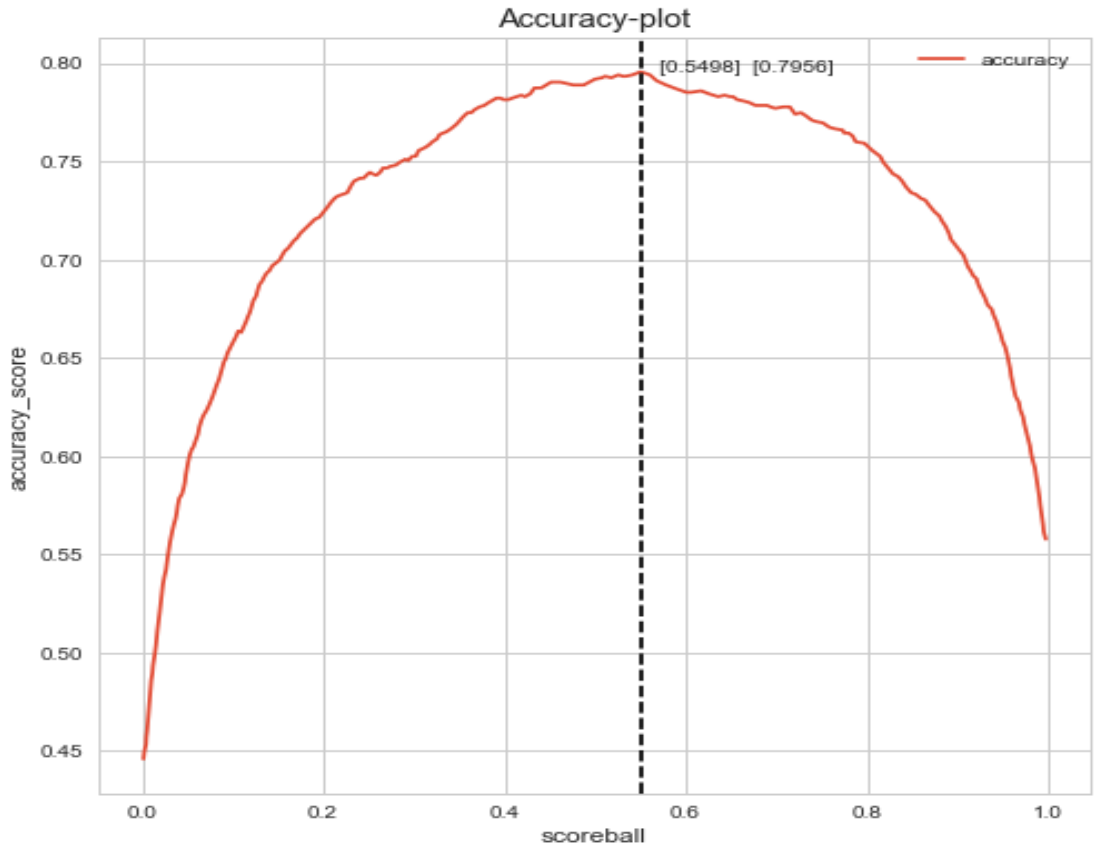

Рисунок 5 – Выбор отсечки

Максимум метрики accuracy на отложенной выборке(one-hold sample) достигается на score= 0.54976983.

Распределение на тестовой выборке

Ниже представлена гистограмма рапспределения скорбаллов на тестовой выборке. Видно, что распределение скорбаллов бимодальное, что лишний раз говорит о неплохом качестве классификации клиентов.

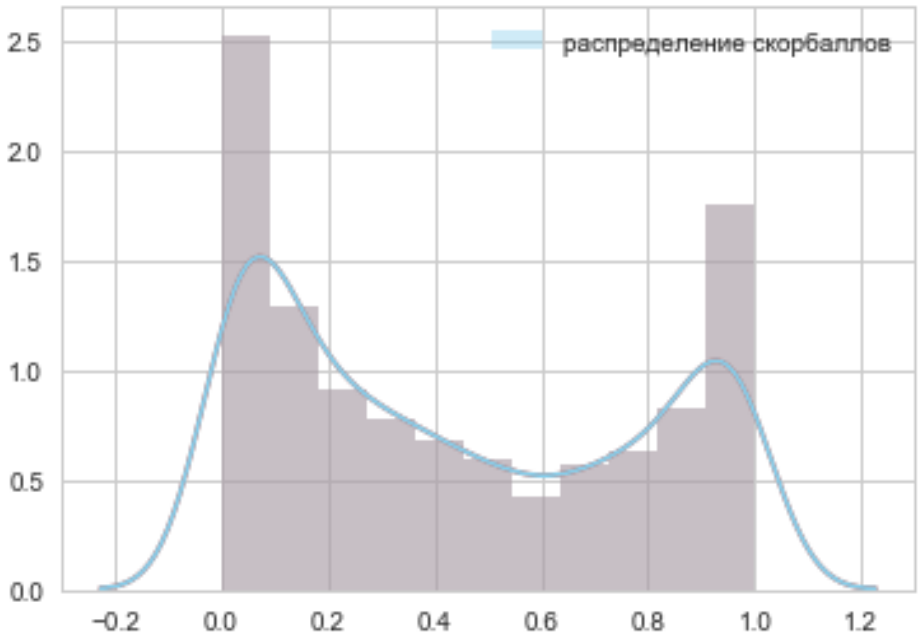

Рисунок 6 – Распределение скорбаллов на тестовой выборке