Решение конкурса BNP Paribas Cardif Claims Management

Попов Артём

ВМК МГУ

Семинары по машинному обучению для 317-ой группы

Постановка задачи

Дана обучающая выборка запросов мененджеров страховой компании BNP Paribas к клиентам (114321 объектов)).

- 131 анонимный признак.
- Из них 19 — категориальные, 112 - вещественные.
- В обучающей выборке 33.7% значений признаков пропущено.

Данные разделены на два класса:

1. Однозначно оплата страховки (24%)
2. Для оплаты нужны дополнительные сведения (76%)

Необходимо классифицировать тестовую выборку (114393 объекта). Метрика качества — log-loss.

Основная идея решения

Один из студентов 317 группы после прошлого соревнования:

«Что-то вы все запарились так...»

Основные этапы решения:

- Строить разные, несложные модели
- Искать хорошие способы для ансамбля

Что такое разные модели:

- Разные алгоритмы машинного обучения
- Разные параметры настройки алгоритмов
- Разные способы работы с категориальными признаками
- Разные способы работы с пропущенными данными

Подробнее о разных моделях

- Разные алгоритмы машинного обучения
  - Градиентный бустинг (XGBoost)
  - Extra trees
  - Random Forest
- Разные параметры настройки алгоритмов
- Разные способы работы с категориальными признаками
  - Ничего не делать
  - Преобразование Бернулли (Наивный Байес)
  - One-hot-encoding
- Разные способы работы с пропущенными данными
  - Заменить на медиану
  - Заменить на -9999 (учитывать как пропущенное значение)

Модели

1. XGBoost на one-hot-encoding
2. XGBoost на one-hot-encoding с большим learning rate
3. Extra trees на one-hot-encoding
4. Random forest на one-hot-encoding
5. XGBoost на one-hot-encoding + преобразование Бернулли
6. Extra trees на one-hot-encoding + преобразование Бернулли
7. Random Forest на one-hot-encoding + преобразование Бернулли
8. XGBoost на отобранных признаках
9. Extra trees на отобранных признаках (Script)
10. Extra trees на отобранных признаках (Script)
11. Extra trees на преобразованной выборке (Script:ExtraNeigbhourLinearFeatures)

Ансамбль

Логистическая регрессия (стэкинг)

- Часть исходных признаков (26 признаков)
- Степени ответов алгоритмов (от 1 до 20)
- Логарифм ответов

Градиентный бустинг (блендинг):

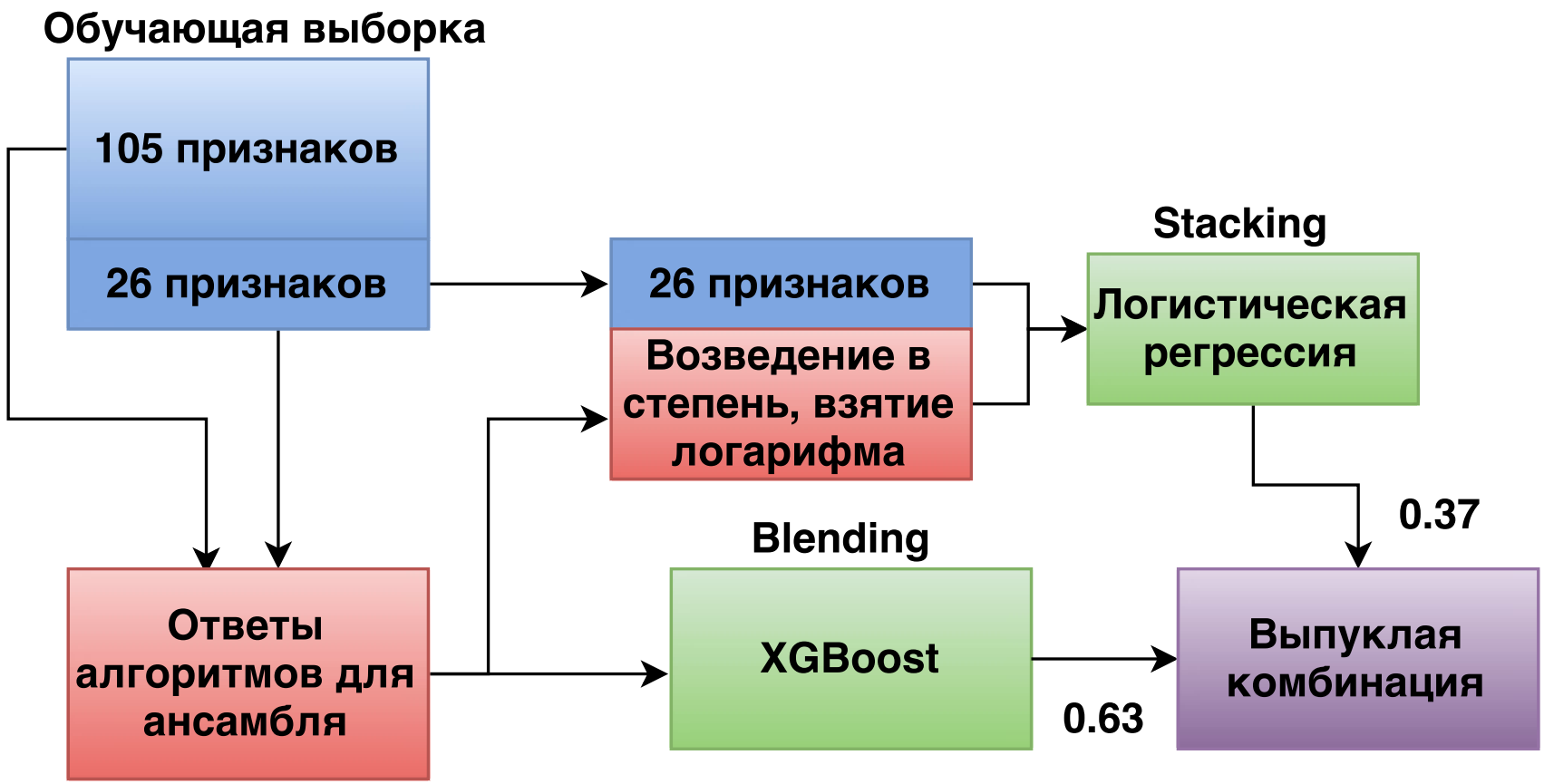
- Ответы алгоритмов
- Калибировка вероятностей

Выпуклая комбинация двух моделей:

- $RES = 0.63 * XGB + 0.37 * LR$

Ансамбль

Схема ансамбля:



Результат

Промежуточные результаты:

- Лучший алгоритм на CV: 0.45318
- Худший алгоритм на CV: 0.46778
- CV: 0.44212

Результат:

- Public LB: 0.44128
- Private LB: 0.43972
- 45 место из 2947 (2 среди 317 группы)

Спасибо за внимание!

Вопросы?