

Логистическая регрессия

1 Задача “Оценка благонадежности заемщика банка”. Есть данные заемщиков банка - физических лиц в файле **credit.txt**. По результатам характера и результата погашений кредита все заемщики разделены на два класса: благонадежный и неблагонадежный (зависимая переменная). Независимыми переменными являются анкетные данные заемщиков:

- Возраст;
- Пол (0 – мужской, 1- женский);
- Семейное положение, состоит ли в браке (да/нет – 1/0);
- Количество иждивенцев;
- Подтвержденный совокупный располагаемый доход;
- Опыт работы, лет;
- Срок проживания в регионе, лет;
- Рыночная стоимость недвижимости в собственности, тыс.долл.;
- Ежемесячный платеж по кредиту, руб.;
- Зависимая переменная (1 – благонадежный, 0 – неблагонадежный заемщик).

Распределение зависимой переменной следующее: 492 благонадежных заемщика из 999.

Разбейте имеющийся датасет на обучающую и тестовую выборки. Обучите модель логистической регрессии и затем для тестовой выборки, считая, что вердикт “благонадежный / неблагонадежный заемщик” не известен, предскажите его. Сравните полученный результат с фактическими данными. Оцените надежность построенной предсказательной модели с использованием кросс-таблицы, метрик и ROC-анализа.

2 Задача “Прогнозирование выживаемости на Титанике”. Используйте уже знакомый Вам [датасет](#) о выживших и погибших пассажирах Титаника. Исходя из логических соображений, отбросьте переменные, которые не могут влиять на возможность пассажира выжить в катастрофе (обоснуйте сделанный выбор). Затем, оставшиеся категориальные переменные (строки) закодируйте целыми числами (например, 0 / 1 для мужчин и женщин). Разбейте полученный датасет

на обучающую и тестовую выборки. Обучите модель логистической регрессии и затем для тестовой выборки, предскажите выжил ли пассажир. Сравните полученный результат с фактическими данными. Оцените надежность построенной предсказательной модели с использованием кросс-таблицы, метрик и ROC-анализа.