# Задание. Прогнозы биномиальной зависимой переменной

Используем данные о займах modeldata::lending\_club. Эти данные были загружены с сайта Lending Club и относятся к первому кварталу 2016 года. Имевшиеся наблюдения и переменные включены только частично. Источник: https://www.lendingclub.com/info/download-data.action

Объясняемая переменная – Class и может быть либо «good» (полностью погашен или гасится без просрочек), либо «bad» (заём списан, дефолт, просрочка от 21 до 120 дней).

funded\_amnt – общая сумма, выделенная по займу к текущему моменту,

term – срок займа в месяцах: «36» или «60»,

int\_rate – процентная ставка в %,

sub\_grade – рейтинг, присвоенный Lending Club,

annual\_inc – указанный заемщиком годовой доход,

emp\_length – стаж работы в годах (возможные значения находятся в диапазоне от 0 до 10, где 0 означает менее одного года, а 10 означает десять или более лет).

Исследуйте характеристики объясняющих переменных. Для funded\_amnt и annual\_inc создайте логарифмы log\_amnt и log\_inc. (Для annual\_inc возьмите доход не менее 1000). Для качественных переменных можно использовать table и barplot. Из term сделайте числовую переменную term\_num в месяцах, из sub\_grade – числовую переменную grade (порядковый номер категории). Из emp\_length сделайте числовую переменную empl=0, 1, 2, …, 9, 10, 11.

Изучите таблицу корреляций для числовых переменных ("bad", "log\_amnt", "term\_num", "int\_rate", "grade", "log\_inc", "empl")

Исследуйте долю плохих случаев в зависимости от term, int\_rate, sub\_grade, emp\_length. Можно использовать table с опцией margin, proportions (prop.table), plot. Процентную ставку можно округлить.

Постройте модель логит (логистическую регрессию) с созданными объясняющими переменными (log\_amnt, term\_num, int\_rate, grade, log\_inc). Команда glm с опцией family = "binomial" (по умолчанию link = "logit" – не требуется указывать)).

Нарисуйте фактические значения и расчетные вероятности (команда fitted) от расчетного индекса (команда predict).

Постройте таблицу сопряженности для предсказанных (с порогом 0.5 и т.д.) и фактических плохих случаев. Постройте кривую ROC (команда pROC::roc, pROC::plot.roc) (?).

Разделите наблюдения на тренировочные и тестовые (20%).

Оцените 4 разных модели: с одной константой, со всеми созданными переменными, и еще какие-то две. Найдите соответствующие прогнозы вероятности predict с опцией type="response".

Постройте кривые ROC для прогнозов вероятности (?).

Сравните прогнозы вероятности по скоринговой функции Брайера (Brier score).