УДК 656.043.4

Использование логистической регрессии для предсказания “проблемных” заказов

А.В. Юдаев

Московский физико-технический институт (государственный университет)

Рассмотрено применение логистической регрессии для классификации поступающих заказов на доставку с целью улучшения качества обслуживания.

Задача стояла в выделении из числа поступающих в режиме реального времени заказов, таких, которые потенциально могли быть доставлены с опозданием. Именно такие заказы будем называть проблемными.

За основу была взята выборка заказов, которая содержала в себе такие данные как: адрес клиента, адрес точки сборки, время и дата оформления заказа, время сборки, время доставки, идентификатор курьера, осуществляющего доставку, сумма заказа, количество позиций (количество товаров в заказе). Помимо выборки заказов, была выборка по курьерам, имевших отношение к доставке заказов. Выборка содержала в себе идентификатор сотрудника, дату и время начала и окончания рабочего дня, отметки времени о нахождении сотрудника на точке сборки с шагом 2 минуты. Также исходные данные были дополнены данными о погоде. В качестве таких данных была выбрана температура воздуха на момент поступления заказа и наличие осадков, с градацией от слабых к сильным. Также проводились тесты с учетом данных о пробках, однако существенной корреляции с итоговым временем доставки обнаружено не было.

На этапе препроцессинга данных необходимо было сбалансировать выборку по заказам, т.к. исходное соотношение проблемных заказов и нормальных было как 2:8. Эта задача решалась простым увеличением весов на записях с проблемными заказами. Также, на этапе препроцессинга, данные по адресам переводились в GPS координаты, после чего вычислялось взаимное расстояние между точками. В итоге записи с адресами клиента и точками сборки заменялись записью об относительном расстоянии между точками. Выборка данных о погоде была в формате METAR, что автоматически позволяло оценивать интенсивность осадков по шкале от 0 до 1. Также стоит отметить, что опытным путем были выбраны еще 2 переменные: количество не готовых к доставке заказов на момент поступления заказа и количество готовых к доставке заказов на момент поступления заказов. Таким образом суммарно эти 2 переменные характеризовали количество заказов в очереди на доставку, на момент поступления нового заказа. После всех преобразований и сопоставлений, данные по каждому фактору (переменные) были нормализованы. Список итоговых факторов представлен в таблице 1.

В качестве классифицирующей модели была выбрана логистическая регрессия с регуляризацией. В качестве параметра регуляризации была выбрана норма L1. Классификатором был фактор времени доставки (1 - опоздали, 0 - доставили вовремя).

Оценка точности модели проходила следующим образом: исходные данные случайным образом разделялись на 2 выборки, в соотношении 7:3. Большая часть данных (70%) использовалась для обучения модели (подбор коэффициентов с минимизацией ошибки), остальная же часть выборки (30%) использовалась для оценки.

Точность модели оценивалась по следующим параметрам: AUC (площадь под кривой на графике True positive - False Positive), Recall и F1. В итоге была достигнута вероятность правильного распознавания 84%. Recall составил 0.537, а F1 составил 0.602. AUC же достиг значения 0.883. График True Positive - False Positive показан на рисунке 1.

Как видно из небольших значений Recall и F1 модель удовлетворительно справляется с задачей классификации при таком наборе факторов. Возможными путями увеличения точности может служить увеличение обучающей выборки, большая натуральная сбалансированность и учет иных факторов (переменных).

Отдельно стоит отметить,что данная модель внедрена в тестовом режиме на работающем предприятии и уже показала свою эффективность, позволив улучшить качество обслуживания клиентов и тем самым увеличить конверсию постоянных пользователей на 10%.

Время заказа	Время оформления заказа. Нормировано в соответствии с 0 -> 10:00, 1 -> 21:40
Признак рабочего дня	1 - рабочий день, 0 - выходной
Сумма заказа	Общая сумма заказа
Количество позиций	Количество позиций в чеке
Количество несобранных	Количество заказов, поступивших перед заказом, но ожидающих сборки.
Количество собранных	Количество заказов, поступивших перед заказом, но еще не закрепленных за курьером.
Количество работающих курьеров	Количество курьеров, активных на момент поступления заказа.
Количество курьеров на точке	Количество курьеров на точке в момент поступления заказа.
Температура воздуха	Температура воздуха на улице в момент оформления заказа.
Осадки	Интенсивность осадков на момент оформления заказа.
Удаленность	Расстояние от точки сборки, до клиента.
Время доставки заказа	0 - время доставки <= 60 минут, 1 - время доставки > 60 минут.

Таблица 1: использованные признаки

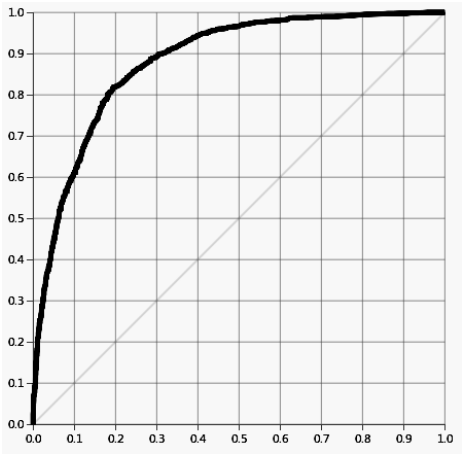

Рис. 1: True Positive - False Positive

Литература

1. *Хардле В.* Прикладная не параметрическая регрессия: Перевод с английского // Москва, Мир, 1993 – 349 с.
2. *Hosmer D. W., Lemeshow S.* Applied Logistic Regression // 2000, John Wiley & Sons, Inc. – 397 с.