Определение мошенничеств в автостраховании на основе метода Cost-sensitive learning

Зеленков Ю.А., д.т.н.
НИУ «Высшая школа экономики»
yzelenkov@hse.ru

Обязательное страхование автогражданской ответственности (ОСАГО) занимает сегодня наиболее значительную долю рынка розничных страховых услуг в России. Однако рентабельность этого вида страховой деятельности в последнее время стремительно сокращается, не в последнюю очередь из-за роста мошенничества. Для борьбы с ним страховые компании вынуждены тщательно проверять все заявки на возмещение убытков, в том числе и те, которые не являются мошенническими. Создание аналитических моделей, которые позволят на основе имеющихся данных о текущих страховых случаях определять попытки мошенничества может сыграть ключевую роль в сокращении подобных затрат.

Проблема определения мошеннических заявок относится к классу задач классификации, которые формулируются следующим образом. Пусть X – множество описаний объектов, а Y – множество наименований классов. Предполагается, что существует функция $f: X \to Y$, значения которой известны на конечной обучающей выборке $X^n = \{(x_1, y_1), \dots, (x_n, y_n)\}$. Требуется построить алгоритм $a: X \to Y$, способный классифицировать произвольный объект $x \in X$. В нашем случае $y_i \in \{0,1\}$, причем значение $y_i = 1$ соответствует мошенническим заявкам.

Наиболее типичной на практике является ситуация, когда количество представителей одного класса в обучающей выборке значительно меньше (в 10 и более раз), чем количество представителей другого класса. В данной работе использовался набор данных, включающий 8517 наблюдений (записей об обращениях за выплатами по ОСАГО), из которых лишь 747 являлись мошенническими. Согласно Abdallah, Maarof & Zainal (2016) эта ситуация является типичной для прогнозирования мошенничеств.

Основные методы классификации несбалансированных данных рассмотрены He & Garcia (2009), к их числу относятся Random Oversampling (дополнение выборки за счет повторения объектов меньшего класса или автоматической генерации объектов, похожих на объекты меньшего класса), Random Undersampling (сокращение объектов большего класса), задание весов классов как метапараметров классификатора и другие.

Для проверки этих методов имеющийся набор данных был разделен на обучающую (6813 объектов, из них 597 мошенничеств) и тестовую выборки (1704 объекта, 150 мошенничеств). На обучающей выборке был обучен ряд методов, включающих различные комбинации способов балансировки набора данных и моделей классификаторов

(логистическая регрессия, случайный лес, градиентный бустинг и т.д.), которые затем были проверены на тестовой выборке. Лучшие результаты по метрике $F_1 = 2(\text{precision} * \text{recall})/(\text{precision} + \text{recall})$ получены для комбинации Random Over Sampling и градиентного бустинга ($F_1 = 0.338$). Соответствующая матрица ошибок представлена в таблице 1, из которой следует, что в данном случае наблюдается большая ошибка в предсказании мошенничеств. Это ведет к затратам на необоснованные выплаты.

Таблица 1. Матрица ошибок. Random Over Sampling и градиентный бустинг.

		Прогноз	
Факт		0	1
	0	1434	120
	1	85	65

Таблица 2. Матрица ошибок. Задание весов классов и градиентный бустинг.

		Прогноз	
Факт		0	1
	0	1299	255
	1	54	96

Аналогичным образом была исследована возможность задания весов классов как метапараметров классификатора, лучшие результаты получены для также для градиентного бустинга ($F_1 = 0.383$, см. таблицу 2). В данном случае количество неверно классифицированных добросовестных заявок на страховые выплаты превышает количество верно предсказанных мошенничеств. Это ведет к значительным затратам на дополнительное рассмотрение заявок.

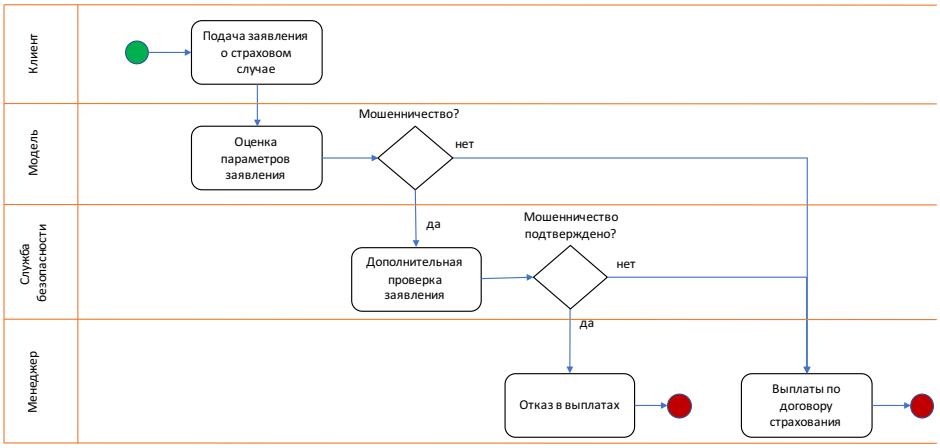

Рис. 1. Бизнес-процесс рассмотрения заявления о страховом случае.

В такой ситуации наиболее эффективным методом является обучение с учетом затрат на реализацию решения классификатора (cost-sensitive learning). Согласно этому методу с каждым элементом матрицы ошибок сопоставляется значение соответствующих затрат и целью обучения является минимизация стоимости принятия решений на основе модели. Для определения затрат рассмотрим бизнес-процесс принятия решения о выплатах по страховому случаю (рис. 1). Матрица затрат в этом случае будет иметь вид,

представленный в таблице 3 (Viane et al., 2007), где – C_A затраты на проверку заявки, C - выплаты по мошенническим заявкам.

Таблица 3. Матрица затрат.

			Факт	
			0	1
Прогноз модели	0	Модель не предсказывает мошенничество. Проверка не производится. Выплаты согласно договору страхования.	0 Выплаты согласно договору. Потерь нет.	C Необоснованные выплаты. Потери страховой компании
	1	Модель предсказывает мошенничество. Проводится проверка заявки. По корректной заявке выплаты производятся, по мошеннической – нет.	C_A	C_A

Исследуем эффективности разных моделей для значений C/C_A в интервале C/C_A = [1,10], который наиболее часто встречается в реальных бизнес - ситуациях. Рассмотрим модели, которые показали лучшие результаты по метрике F_1 без cost-sensitive обучения: ROS-XGB (Random Oversampling + XGBoost); CW-XGB (Class Weight + XGBoost); ROS-LR (Random Oversampling + Logistic Regression); CW-LR (Class Weight + Logistic Regression). В качестве метрики качества будем использовать показатель сокращения затрат (Bahnsen et al., 2014):

$$Savings\big(a(S)\big) = \frac{Cost_l(S) - Cost\big(a(S)\big)}{Cost_l(S)},$$

где $Cost_l(S) = \min\{Cost(0), Cost(1)\}$ – минимальные затраты при отнесении всех объектов к одному классу (0 или 1), $a(S)$ - прогноз, S - тестовая выборка.

Рис.2. Сокращение затрат при использовании различных моделей

Зависимость $Savings$ от отношения C/C_A для различных моделей показана ни рис. 2, из которого следует, что выбор модели зависит от конкретного соотношения затрат.

Из представленных результатов можно сделать вывод, что cost-sensitive learning является предпочтительным методом обучения в случае несбалансированных данных, поскольку он позволяет связать решения классификатора с реальными затратами бизнеса. Отметим, однако, что задание одной матрицы затрат для всех наблюдений (таблица 3) может также вести к неэффективным моделям, строго говоря, затраты различны для разных наблюдений (Bahnsen et al., 2014; Bahnsen, Aouada & Ottersten, 2015), но для решения задачи в такой постановке необходимы дополнительные данные об объеме выплат по каждому страховому случаю.

Литература

1. Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications*, 68, 90-113.
2. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9), 1263-1284.
3. Viaene, S., Ayuso, M., Guillen, M., Van Gheel, D., & Dedene, G. (2007). Strategies for detecting fraudulent claims in the automobile insurance industry. *European Journal of Operational Research*, 176(1), 565-583.
4. Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, 42, 6609-6619.
5. Bahnsen, A. C., Stojanovic, A., Aouada, D., & Ottersten, B. (2014). Improving credit card fraud detection with calibrated probabilities. In: *Proceedings of the 2014 SIAM International Conference on Data Mining* (pp. 677-685). Society for Industrial and Applied Mathematics.