Моделювання в економіці, організація виробництва та управління проектами

УДК 386.01:004.89

К.А. Базилевич, М.С. Мазорчук, А.А. Сухобрус

Национальный аэрокосмический университет имени Н.Е. Жуковского «ХАИ», Харьков

ОПРЕДЕЛЕНИЕ ВЕРОЯТНОСТИ ВЫПЛАТ ПО СТРАХОВЫМ ПОЛИСАМ НА ОСНОВЕ МЕТОДОВ DATA MINING

На вероятность выплат по страховым полисам большое влияние оказывают социально-экономические характеристики страхователей. Для определения значимых факторов и оценивания степени влияния чаще всего используют статистические методы анализа данных. Однако, в последнее время широкое применение получили методы Data Mining (интеллектуального анализа данных), которые при больших объемах информации и сложных взаимосвязях могут давать более точные оценки. Для решения задачи оценивания вероятности страховых выплат в данной работе предлагается использовать логистическую регрессию и байесовский классификатор. Для нахождения параметров логистической регрессии предлагается использовать различные методы в зависимости от вида исходных данных.

Ключевые слова: *личное страхование, оценка вероятности, логистическая регрессия, байесовский классификатор, метод оценки шансов, метод максимального правдоподобия.*

Введение

Эволюция рыночных отношений требует поиска новых подходов для управления финансами страховых компаний, что обусловлено необходимостью обработки больших объемов данных и применения сложных математических расчетов. В современных реалиях невозможно управлять объемными финансовыми потоками страховой компании без применения информационных технологий, без возможности прогнозирования и анализа данных. В сложившейся экономической ситуации количество рисков для страховщиков постоянно растет. Существует острая необходимость в инструментах регулирования и анализа финансовых потоков в страховых фондах, а также факторов, влияющих на этот процесс.

Теоретическому обоснованию целесообразности применения инструментов страхования посвящено множество отечественных и зарубежных работ [1 – 7]. Среди авторов научных работ в данном направлении можно перечислить В.Д. Базилевича, О.А. Гаманкова, Н.Н. Внукова, А.Д. Зарубы, С.С. Осадец, А.Г. Шоломицкого, А.О. Недосекина, V. Malinovskii, C.D. Daykin и др. На данный момент разработаны теоретические основы страхования [3], большое внимание уделяется актуарным методам [5], разработаны методики расчета тарифных ставок по отдельным видам страхования [7].

Следует указать, что методы статистического анализа, которые используют отечественные страховщики для оценки будущей прибыли и величины страховых выплат, не всегда позволяют с достаточной степенью достоверности оценить будущие затраты и расходы по страховым взносам и выплатам. Аналитические расчеты, которые часто используют актуарии, также не всегда дают верные результаты в условиях постоянных изменений внешней среды, что подробно описано в работе [8].

Чаще всего страховой договор предусматривает рассрочку взносов, что представляет еще один риск для страховщиков – риск прекращения выплат страховых премий раньше установленного срока, что приводит к убыточности отдельного страхового договора, а при большом количестве таких «неплательщиков» – и к убыточности всего страхового фонда.

Поэтому, актуальным является рассмотрение новых методов и моделей анализа данных, позволяющих уменьшить риски страховщиков и изначально классифицировать страхователей с разными социально-экономическими характеристиками на классы по платежеспособности.

Постановка задачи

На платежеспособность отдельного страхователя влияет целый ряд факторов, таких как: средний уровень дохода, место жительства, наличие банковского счета, количество детей, наличие постоянной работы и т.п. Задача исследования состоит в определении и анализе моделей и методов, которые позволяют оценить вероятность отсутствия денежных поступлений (премий) по договорам страхования от определенного страхователя с известными индивидуальными характеристиками.

С математической точки зрения эту задачу можно отнести к задачам классификации «с учителем»: необходимо найти функциональную зависимость, которая позволит разделить страхователей на классы «Плательщиков» и «Неплательщиков» на основе данных выборочной (обучающей) популяции. Для такого класса задач целесообразно использование пороговой дискриминативной функции – логистической регрессии, которая широко применяется для определения вероятностей возникновения некоторого события при заданных значениях множества характеристик [9].

В статье предлагаются методы для оценивания параметров логистической регрессии для разных условий. В случае одной политомической входной переменной с минимальным числом категорий – метод оценки шансов и вероятностей. В данном случае качество классификации можно оценивать для каждой входной переменной отдельно: оценка не зависит от связанности входных переменных, что позволяет не проводить проверку коррелированности и предварительный отбор значимых переменных [10]. Для нескольких объясняющих переменных предлагается использовать байесовский классификатор, который позволит при отсутствии корреляции между признаками отнести конкретных индивидов рассматриваемой популяции к некоторому классу по платежеспособности. При наличии корреляции факторных признаков и сложных зависимостей между входными переменными предлагается использовать известный метод максимального правдоподобия.

В результате анализа страховщик может иметь готовый математический аппарат, позволяющий на практике получить значения вероятностей по выплатам в страховых полисах страхователей в условиях различных исходных данных.

Оценивания параметров логистической регрессии на основе метода оценки шансов и вероятностей

Рассмотрим некоторую выборку страхователей на основании данных из источника [11]. По каждому страхователю известна информация о выплатах по страховому полису. Объясняющей переменной в данном случае является уровень дохода страхователя. Данная переменная является политомической. Каждый страхователь может принадлежать по уровню дохода к трем классам «низкий уровень доходов», «высокий уровень доходов» и «средний уровень доходов». Также рассматриваются два события: выплата по страховому полису была произведена (y=0) и отсутствовала (y=1).

Необходимо оценить параметры логистического уравнения для данной задачи и определить, с какой вероятностью страхователь не произведет выплату, т.е. оценить его платежеспособность.

Вероятность того, что выходная переменная $y = 1$ для заданного значения объясняющей переменной x будет $P(y = 1 | x) = \rho(x)$, а вероятность того, что $y = 0$ при заданном значении x будет равна $P(y = 0 | x) = 1 - \rho(x)$.

Условное среднее для логистической регрессии в данном случае определяется как:

$$\rho(x) = e^{g(x)} \Big/ \left(1 + e^{g(x)}\right),$$

где $g(x) = \beta_0 + \beta_1 C_1 + \beta_2 C_2$, C_1, C_2 – переменные для квантования значений в трех интервалах; x – объясняющая переменная; $\beta_0, \beta_1, \beta_2$ – искомые параметры, $c(x)$ – вероятность события. Функция определена на бесконечном интервале и принимает значения в диапазоне [0, 1]. Требуется найти наилучшие оценки параметров $\beta_0, \beta_1, \beta_2$. Упорядочим информацию по страхователям на основе данных [11] в виде таблицы (табл. 1).

Таблица 1

Данные о страхователях по уровню доходов

Исход	Низкий уровень доходов	Высокий уровень доходов	Средний уровень доходов	Всего
y=0	21	52	62	135
y=1	7	1	7	15
Всего	28	53	69	150

В табл. 1 в строке с классом «низкий уровень доходов» переменные квантования будут равны: $C_1 = C_2 = 0$. В строке с классом «высокий уровень доходов» $C_1 = 1$, $C_2 = 0$. В строке с классом «средний уровень доходов» $C_1 = C_2 = 1$.

Шансы не произвести выплату для всех категорий уровней доходов оценивается по формулам:

$$Ch_{y=1,C_1} = \frac{7}{21} \approx 0.33,\ Ch_{y=1,C_2} = \frac{1}{52} \approx 0.02,$$

$$Ch_{y=1,C_3} = \frac{7}{62} \approx 0.11.$$

Отношение шансов для категорий «высокий уровень доходов» к категории «низкий уровень доходов» оценивается по формуле:

$$OR\left(\frac{C_2}{C_1}\right) = \frac{Ch_{y=1,C_2}}{Ch_{y=1,C_1}} = 0.057.$$

Отношение шансов для категорий «средний уровень доходов» к категории «низкий уровень доходов» оценивается по формуле:

$$OR\left(\frac{C_3}{C_1}\right) = \frac{Ch_{y=1,C_3}}{Ch_{y=1,C_1}} = 0.339.$$

Экспериментальную вероятность отсутствия выплаты для категории «низкий уровень доходов» можно найти, поделив число положительных исходов на общее количество исходов

$$\rho_{exp} = 7/28 = 0.25,$$

отсюда коэффициент β_0 может быть найден как

$$\beta_0 = \ln\left(\frac{c_{exp}}{1-c_{exp}}\right) = -1.099.$$

Для категории «высокий уровень доходов» экспериментальную вероятность отсутствия выплаты можно оценить по формуле

$$c_{exp} = \frac{1}{53} = 0,019.$$

отсюда коэффициент β_1 может быть найден как

$$\beta_1 = \ln\left(\frac{c_{exp}}{1-c_{exp}}\right) - \beta_0 = -2.85.$$

Для категории «средний уровень доходов» экспериментальную вероятность отсутствия выплаты можно оценить по формуле

$$c_{exp} = \frac{7}{69} = 0.101.$$

отсюда коэффициент β_2 может быть найден как

$$\beta_2 = \ln\left(\frac{c_{exp}}{1-c_{exp}}\right) - \beta_0 - \beta_1 = 1.77.$$

Вероятность того, что выходная переменная y будет равна единице (т.е. выплата по страховому полису не будет произведена) для категории «Низкий уровень доходов» рассчитывается по формуле

$$P(y=1|x) = \frac{e^{\beta_0}}{1+e^{\beta_0}} = \frac{e^{-1.099}}{1+e^{-1.099}} \approx 0.25.$$

Вероятность того, что выходная переменная $y=1$ для категории «Высокий уровень доходов» рассчитывается по формуле

$$P(y=1|x) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} = \frac{e^{-1.099-2.85}}{1+e^{-1.099-2.85}} \approx 0.019.$$

Вероятность того, что выходная переменная $y=1$ для категории «Средний уровень доходов» рассчитывается по формуле

$$P(y=1|x) =$$

$$= \frac{e^{\beta_0+\beta_1+\beta_2}}{1+e^{\beta_0+\beta_1+\beta_2}} = \frac{e^{-1.099-2.85+1.77}}{1+e^{-1.099-2.85+1.77}} \approx 0.101.$$

Можно сделать вывод о том, что чем выше уровень доходов страхователя, тем меньше вероятность того, что данный страхователь перестанет вносить выплаты по страховому полису.

Оценка вероятности выплат по страховым полисам с использованием байесовского классификатора

Рассмотрим выборку из 30 страхователей входными переменными, заданными в номинальной шкале (табл. 2) на основании данных из источника [11]. Для анализа используем следующие признаки: возраст (в годах), срок проживания в данной местности, владение недвижимостью и банковским счетом. По данным табл. 2 были рассчитаны коэффициенты парной корреляции, значения которых находятся в интервале [-0.276; 0.067], что говорит о низкой корреляционной зависимости между входными переменными.

Таблица 2

Данные о страхователях
по личным характеристикам

№	Возраст	A1	A2	A3	A4
1	40-50	<40	нет	нет	да
2	50-60	<40	да	нет	нет
3	50-60	<40	да	нет	нет
4	40-50	<40	да	нет	нет
5	40-50	<40	нет	да	да
6	50-60	>40	нет	нет	да
7	40-50	<40	да	да	да
8	50-60	<40	нет	да	да
9	40-50	<40	да	нет	да
10	50-60	<40	да	нет	да
11	50-60	<40	нет	нет	нет
12	40-50	<40	да	да	да
13	40-50	<40	да	нет	да
14	50-60	<40	нет	да	да
15	40-50	>40	нет	нет	нет
16	50-60	>40	нет	нет	да
17	50-60	<40	нет	нет	да
18	40-50	<40	нет	нет	да
19	40-50	<40	нет	да	да
20	50-60	<40	нет	нет	да
21	50-60	<40	нет	нет	да
22	40-50	>40	да	нет	да
23	40-50	<40	нет	нет	да
24	40-50	<40	нет	нет	нет
25	50-60	<40	нет	да	да
26	40-50	<40	да	нет	да
27	40-50	<40	нет	да	нет
28	40-50	>40	да	нет	нет
29	50-60	>40	нет	нет	да
30	40-50	<40	нет	нет	да

*Примечание. **A1** – срок проживания в данной местности, лет; **A2** – владеет банковским счетом; **A3** – владеет недвижимостью; **A4** – выплата по договору*

Таким образом, в данном случае можно использовать байесовский классификатор, применение которого для этого случая подробно рассмотрено в работе [10].

Обозначим как C_1 класс страхователей «Неплательщики», для которых выплата не производится (значение результирующей переменной – «нет»). Через C_2 можно обозначить класс страхователей «Плательщики», который вносят плату за страхование (значение результирующей переменной – «да»).

Применение байесовского классификатора не дает возможности получить вид статистической зависимости на основе обучающей выборки, однако позволяет определить вероятность того, что страхо-

ватель с заданными характеристиками попадет в тот или иной класс. Например, определим, что страховщик возраста от 40 до 50 лет, проживающий в данной местности менее 40 лет и не владеющий банковским счетом и недвижимостью попадет в класс «Неплательщики».

Необходимо максимизировать произведение вероятностей $P(X \mid C_k)P(C_k)$ для $k = 2$, т.к. в данной задаче всего два класса. Априорная вероятность появления класса C_1 вычисляется по формуле $P(C_1) = \frac{8}{30} = 0.27$, априорная вероятность появление класса C_2 вычисляется по формуле

$$P(C_2) = 22/30 = 0.73.$$

Всего наблюдаемых примеров 30, 22 из них - «Плательщики», 8 – «Неплательщики».

Условные вероятности для определения $P(X \mid C_k)$ рассчитані в табл. 3. Рассчитаем обобщенные вероятности $P(X \mid C_k)$ для событий:

$$P(X \mid C_1) = 0.625 \cdot 0.750 \cdot 0.500 \cdot 0.875 = 0.205,$$
$$P(X \mid C_2) = 0.545 \cdot 0.818 \cdot 0.682 \cdot 0.682 = 0.207.$$

Тогда вероятности $P(X \mid C_k)P(C_k)$ будут соответственно равны:

$$P(X \mid C_1)P(C_1) = 0.205 \cdot 0.73 = 0.150,$$
$$P(X \mid C_2)P(C_2) = 0.207 \cdot 0.27 = 0.055.$$

Таблица 3

Условные вероятности для данных о страхователях

Описание вероятности	Расчет
P(возраст 40 – 50 | C_2)	12 / 22 = 0.545
P(возраст 40 – 50 | C_1)	5 / 8 = 0.625
P(проживает < 40 | C_2)	18 / 22 = 0.818
P(проживает < 40 | C_1)	6 / 8 = 0.750
P(нет банковского счета | C_2)	15 / 22 = 0.682
P(нет банковского счета | C_1)	4 / 8 = 0.500
P(нет недвижимости | C_2)	15 / 22 = 0.682
P(нет недвижимости | C_1)	15 / 7 = 0.857

Выбирается тот класс, вероятность для которого больше, т.е. рассматриваемый страхователь относится к классу C_1 – «Неплательщики».

Нормализация вероятностей может выглядеть следующим образом:

$$P'(X \mid C_1)P(C_1) = \frac{0.150}{0.150 + 0.055} = 0.731,$$

$$P'(X \mid C_2)P(C_2) = \frac{0.150}{0.055 + 0.150} = 0.269.$$

Таким образом, страхователь с описанными характеристиками с вероятностью 0.269 окажется платежеспособным (попадет в класс «Плательщики»), а с вероятностью 0.731 окажется неплатежеспособным (попадет в класс «Неплательщики»).

Оценивания параметров логистической регрессии на основе метода максимального правдоподобия

Рассмотрим выборку из 150 страхователей с входными характеристиками, представленными на рис. 1 на основании данных из источника [11]. Результирующий признак измеряется в дихотомической шкале, а факторные признаки – в метрических и других видах шкал. Необходимо определить вероятность платежеспособности страхователя с данным множеством характеристик.

Поскольку метод максимального правдоподобия для определения параметров логистической регрессии является достаточно трудоемким, то для демонстрации работы данного метода используем известный программный статистический инструментарий обработки данных SPSS [12].

Реализованный метод оценки параметров логистической регрессии в SPSS позволяет не только определить параметры модели и оценить вероятность, но и проанализировать качество построенной модели, что является важным при оценивании страхователей. Наиболее значимые результаты расчетов представлены в таблицах ниже. В частности, в табл. 4 представлены коэффициенты качества модели.

Таблица 4

Сводная таблица модели

Шаг	-2 Log правдоподобие	R-квадрат Кокса и Снелла	R-квадрат Нейджелкерка
1	28.135	0.370	0.775

-2 Log правдоподобие – эта величина, которая характеризует соответствие модели исходным данным. Чем меньше значение данного показателя, тем адекватнее сформирована модель. Остальные критерии, приведённые в табл. 4, устойчивее традиционных статистик согласия, используемых в логистической регрессии, особенно для моделей с непрерывными ковариатами и для исследования выборок малого объема. В мере определенности по Коксу и Снеллу значение равное единице является теоретически достижимым. Этот недостаток устранен благодаря модификации данной меры по методу Нейджелкерка.

Данные критерии показывают долю влияния всех факторных признаков на дисперсию зависимой переменной. Чем ближе коэффициенты к единице, тем лучше.

Подробную информацию о методике расчета данных критериев и анализе их числовых характеристик можно найти в источниках [13 – 15].

В табл. 5 представлены значения критерия Хи-квадрат.

	Возраст	Пол	Срок_проживания_в_данной_местности	Профессия_риск	Место_работы	Владеет_банковскими_счетами	Работает_на_предприятии	Владеет_недвижимостью	Выплата_в_срок	Ожидаемые	Группа	Остатки	п
1	46	ж	32	Другое	Другое	Нет	17	Нет	Да	1,00000	Да	,00000	
2	55	м	30	Высокий риск	Гос. учреждение	Да	11	Нет	Да	1,00000	Да	,00000	
3	51	ж	4	Низкий риск	Другое	Да	0	Нет	Нет	,99702	Да	-,99702	
4	44	м	20	Низкий риск	Гос. учреждение	Да	6	Нет	Да	1,00000	Да	,00000	
5	48	м	4	Низкий риск	Гос. учреждение	Нет	1	Да	Да	1,00000	Да	,00000	
6	51	ж	48	Другое	Гос. учреждение	Нет	26	Нет	Да	1,00000	Да	,00000	
7	43	ж	29	Другое	Гос. учреждение	Да	13	Да	Да	1,00000	Да	,00000	
8	53	м	34	Низкий риск	Другое	Нет	33	Да	Да	1,00000	Да	,00000	
9	49	ж	17	Низкий риск	Другое	Да	0	Нет	Да	,99857	Да	,00143	
10	59	м	30	Другое	Другое	Да	7	Нет	Да	,99990	Да	,00010	
11	59	м	51	Другое	Гос. учреждение	Нет	15	Нет	Да	1,00000	Да	,00000	
12	45	м	28	Другое	Другое	Да	13	Да	Да	1,00000	Да	,00000	
13	49	м	34	Высокий риск	Другое	Да	23	Нет	Да	1,00000	Да	,00000	
14	50	ж	28	Высокий риск	Другое	Нет	21	Да	Да	1,00000	Да	,00000	
15	48	м	44	Высокий риск	Другое	Нет	3	Нет	Нет	,07367	Нет	-,07367	
16	58	м	49	Низкий риск	Другое	Нет	33	Нет	Да	1,00000	Да	,00000	
17	53	ж	9	Низкий риск	Гос. учреждение	Нет	8	Нет	Да	1,00000	Да	,00000	
18	42	м	32	Другое	Другое	Нет	7	Нет	Да	,97958	Да	,02042	
19	44	м	9	Низкий риск	Гос. учреждение	Нет	3	Да	Да	1,00000	Да	,00000	
20	54	м	40	Низкий риск	Другое	Нет	23	Нет	Да	1,00000	Да	,00000	
21	57	м	24	Низкий риск	Другое	Нет	19	Нет	Да	1,00000	Да	,00000	
22	48	м	47	Другое	Гос. учреждение	Да	29	Нет	Да	1,00000	Да	,00000	

Рис. 1. Фрагмент данных о выборке страхователей в файле SPSS

Таблица 5

Универсальный критерий коэффициентов модели

Шаг		Хи-квадрат	Степени свободы	Значимость
1	Шаг	69.390	9	0.000
	Блок	69.390	9	0.000
	Модель	69.390	9	0.000

В табл. 6, 7 представлена классификация случаев по критерию Хосмера - Лемешова. Часть дисперсии, объяснимой с помощью логистической регрессии, в нашем случае составляет 48,4%, т.е. в рассматриваемой задаче значение р-уровня равно 0.484, что говорит о высокой степени согласованности модели (считается, что при р> 0,05 модель адекватно описывает данные).

По сути, критерий Хосмера-Лемешова – это оценка согласия модели с реально существующими в выборке частотами [16 – 18].

Этот критерий отражает наличие в модели факторов, представляющих собой «информационный мусор», который приводит к снижению качества модели и снижению значения по этому критерию.

Таблица сопряженности для проверки согласия Хосмера-Лемешева (см. табл. 7) строится таким образом: на основе расчетного значения вероятностей зависимой переменной рассчитывают децили, которые разделяют значение вероятности платежеспособности на 10 групп. Далее строят таблицу связанности, строки которой задают группы децилей риска, а столбцы – зависимую бинарную переменную «Выплата в срок».

На основе критерия согласия Хи-квадрат сравнивают степень различий фактических и ожидаемых частот в полученной таблице связанности.

Таблица 6

Критерий Хосмера-Лемешова

Шаг	Хи-квадрат	Степени свободы	Значимость
1	7.497	8	0.484

Таблица 7

Таблица сопряженности для проверки согласия Хосмера-Лемешева

		Выплата = Нет		Выплата = Да		Всего
		Наблюденные	Ожидаемые	Наблюденные	Ожидаемые	
Шаг 1	1	14	11.749	1	3.251	15
	2	0	2.954	15	12.046	15
	3	1	0.286	14	14.714	15
	4	0	0.011	15	14.989	15
	5	0	0.000	15	15.000	15
	6	0	0.000	15	15.000	15
	7	0	0.000	15	15.000	15
	8	0	0.000	15	15.000	15
	9	0	0.000	15	15.000	15
	10	0	0.000	15	15.000	15

В табл. 8 приведены проценты, отображающие разные уровни классификации модели. Получены достаточно высокие показатели, т.е. более 90% случаев удалось классифицировать верно.

Таблица 8

Таблица классификации

Наблюдаемые			Предсказанные		
			Выплата в срок		Процент правильных
			Нет	Да	
Шаг1	Выплата в срок	Нет	14	1	93.3
		Да	1	134	99.3
	Общая проц. доля				98.7

В табл. 9 приведены параметры уравнения логистической регрессии.

Таблица 9

Переменные уравнения регрессии

Влияющая переменная	Коэф. уравнения регрессии β	Среднекв. ошибка	Статистика Вальда	Уровень значимости
A – «Возраст»	0.106	0.110	0.931	0.335
B – «Пол»	2.604	1.593	2.671	0.102
C – «Срок проживания в данной местности»	0.073	0.060	1.476	0.222
D – «Профессия»			8.889	0.012
D1 – «Профессия с низким уровнем риска»	3.954	1.703	5.391	0.020
D2 – «Профессия с высоким уровнем риска»	-3.986	1.784	4.991	0.003
E – «Место работы»	0.990	.361	7.509	0.025
F – «Срок работы на предприятии»	-3.693	1.555	5.643	0.006
G – «Владеет банковским счетом»	-20.089	5644.593	0	0.018
H – «Владеет недвижимостью»	13.963	5644,596	0	0.997
I – «Константа»	-3.986	1.784	4.991	0.998

На основе данной таблицы можно определить наиболее значимые факторы, по которым можно получить наименьшие ошибки с высокой долей вероятности.

Общий вид уравнения регрессии для страхователя будет иметь вид:

$$g(x) = 0.106 \cdot A + 2.604 \cdot B + 0.073 \cdot C + Risk - E \cdot 3.986 + \\ + 0.99 \cdot F - 3.693 \cdot G - 20.089 \cdot H + 13.963,$$

где $Risk = \begin{cases} 3.954, \text{ если } D = D1, \\ 7.607, \text{ если } D = D2. \end{cases}$

Например, для страхователя возраста 25 лет, который проживает в данной местности в течение 5 лет, имеет один год опыта работы, владеет банковским счетом и владеет недвижимостью, будет справедливо выражение:

$$g(x) = 0.106 \cdot 25 + 2.604 + 0.073 \cdot 5 + 7.607 - \\ - 3.986 + 0.99 \cdot 1 - 3.693 - 20.089 + 13.963 = 0.411.$$

Тогда вероятность того, что такой страхователь окажется платежеспособным, вычисляется по формуле

$$c(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \approx 0.6.$$

При этом, как видно из табл. 9, по статистике Вальда наиболее значимыми являются такие факторы: профессия, связанная с рисками (значение 8.889), трудовой стаж (значение 7.509), наличие банковского счета (значение 5.643).

Тест Вальда позволяет оценить справедливость линейной регрессионной модели. Этот тест является более приемлемым по сравнению с тестами отношения правдоподобия и множителей Лагранжа, которые используются используемый для проверки ограничений на параметры статистических моделей, оценённых на основе выборочных данных [19 – 20]. Основным достоинством данного теста является то, что доверительный интервал теста представляет собой замкнутую форму. Чем выше значение статистики Вальда, тем лучше.

Также значимость факторов подтверждается соответствующем уровнем значимости. Значимость определяется как р-уровень – рассчитанная в ходе статистического теста вероятность ошибочного отклонения нулевой гипотезы. Чем меньше р-уровень, тем более значимой называется тестовая статистика.

На основе исходных данных (см. рис. 1) были рассчитаны вероятности попасть в группу «Плательщики» для всех наблюдений анализируемой совокупности.

В столбцах «Ожидаемые» и «Группа» приведены ожидаемые вероятности попадание страхователей в группу «Плательщик» (значение «Да») или «Неплательщик» (значение «Нет»).

В столбце «Остатки» приведены значения разницы между ожидаемыми значениями вероятности и наблюдаемыми.

Как видно из результатов, построенная модель действительно адекватно описывает данную совокупность.

Выводы

Таким образом, в данной работе проанализированы существующие методы оценки параметров логистической регрессии, используемые для оценивания платежеспособности страхователей и убыточности страховых договоров. Показано, в каких случаях целесообразно применение тех или иных методов для определения вероятности и оценивания параметров моделей. Данные модели не являются статичными. Расчет параметров можно проводить каждый раз при изменении клиентской базы страхового фонда, а использование программного инструментария SPSS позволит осуществлять расчеты достаточно оперативно. Полученные данные позволят более точно оценивать страхователей и анализировать

риски страховых компаний в условиях постоянно меняющейся рыночной среды.

Список литературы

1. Страхування [Текст]: підручник / В.Д. Базилевич; за ред. В.Д. Базилевича. – К.: Знання, 2008. – С. 687-690.

2. Рейтинг страхових компаній [Електронний ресурс]: за даними страхового журналу «Forinsurer». Режим доступа: http://forinsurer.com/files/file00556.pdf - 20.11.2015 г.

3. Сахирова Н.П. Страхование [Текст]: учеб. пособие / Н.П. Сахирова. – М.: ТК Велби, 2006. – 724 с.

4. Александрова М.М. Страхування [Текст]: учбово-методичний посібник / М.М. Александрова. – К.: ЦУЛ, 2002. – С. 5-30.

5. Фалин Г.И. Актуарная математика в задачах [Текст] / Г.И. Фалин, А.И. Фалин. – М.: ФИЗМАТЛИТ, 2003. – С. 81-129.

6. Шоломицкий А.Г. Финансирование накопительных пенсий: актуарные методы и динамические модели [Текст]: А.Г. Шоломицкий // Обозрение прикладной и промышленной математики. – 2002. – Т. 9. – С. 544-577.

7. Pomazkin D. The Method of Scenario Analysis based on Space of Solutions [Text] / D. Pomazkin // ACTUARY Information and analytical bulletin. – 2010-11. – № 1 (4). – P. 72-73.

8. Гужва В.М. Интеллектуальные системы поддержки принятия решений в страховании: потребности украинских страховых компаний и их удовлетворение [Текст] / В.М. Гужва, А.С. Скрипова // Бизнес Информ. – 2012. – №3. – С. 183-187.

9. Диденко Н.И. Методы анализа процессов в мировой экономике: Методические указания [Текст] / Н.И. Диденко. – СПб.: Санкт-Петербургский Государственный Политехнический Университет, 2007. – 25 с.

10. Паклин Н.Б. Бизнес-аналитика: от данных к знаниям [Текст]: учеб. пособие / Н.Б. Паклин, В.И. Орешков. – СПб.: Питер, 2013. – С. 342-423.

11. Выборка данных о клиентах банка и страховой компании [Электронный ресурс]: по данным Московского финансового университета. Режим доступа к ресурсу: https://www.google.com.ua/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwjesPPHrejJAhUnc3IKHSXmBe0QFggbMAA&url=http%3A%2F%2Ffa-kit.ru%2Fcommon%2FITBank%2FScoring.xls&usg=AFQjCNHLORJTx-LWM87crAMwIO6nbEPd0w&sig2=_iCwvHI3HIF68QrPzgxiWA&cad=rja - 10.12.2015 г.

12. Пациорковский В.В. SPSS для социологов [Текст]: учеб. пособие / В.В. Пациорковский, В.В. Пациорковская. – М.: ИСЭПН РАН, 2005. – 433 с.

13. Allison Paul. What's the Best R-Squared for Logistic Regression? [Electronic resource] / AccessMode: http://statisticalhorizons.com/r2logistic - 08.01.2016.

14. Cox D.R. Analysis of Binary Data [Text] / D.R.Cox, E.J. Snell.– Chapman and Hall/CRC, 1989. – 240 p.

15. Nagelkerke N.J.D. Note on a General Definition of the Coefficient of Determination [Text] / N.J.D. Nagelkerke // Biometrika. – 1991. – Vol. 78, №3. – P. 691-692.

16. Hosmer-Lemeshow Statistic [Electronic resource] / Access Mode: https://www.biostat.wisc.edu/~cook/642.tex/notes0412.pdf - 08.01.2016.

17. Hosmer-Lemeshow Test [Electronic resource] / Access Mode: http://www.real-statistics.com/logistic-regression/hosmer-lemeshow-test/ - 08.01.2016.

18. Bartlett Jonathan. The Hosmer-Lemeshow goodness of fit test for logistic regression [Electronic resource] / Jonathan Bartlett. – Access Mode: http://thestatsgeek.com/2014/02/16/the-hosmer-lemeshow-goodness-of-fit-test-for-logistic-regression/ - 08.01.2016.

19. Tests of Hypotheses [Electronic resource] / Access Mode:http://data.princeton.edu/wws509/notes/c2s3.html- 08.01.2016.

20. Магнус Я.Р. Эконометрика. Начальный курс. [Текст]: учеб. / Я.Р. Магнус, П.К.Катышев, А.А. Пересецкий – М.: Дело, 2004. – 576 с.

Поступила в редколлегию 20.01.2016

Рецензент: д-р техн. наук Р.М. Трищ, Украинская инженерно-педагогическая академия, Харьков.

ВИЗНАЧЕННЯ ЙМОВІРНОСТІ ВИПЛАТ ЗА СТРАХОВИМИ ПОЛІСАМИ НА ОСНОВІ МЕТОДІВ DATA MINING

К.О. Базілевич, М.С. Мазорчук, А.А. Сухобрус

На ймовірність виплат за страховими полісами мають великий вплив соціально-економічні характеристики страхувальників. Для визначення значущих чинників і оцінювання ступеня впливу найчастіше використовують статистичні методи аналізу даних. Однак, останнім часом широке застосування отримали методи Data Mining (інтелектуального аналізу даних), які при великих обсягах інформації та складних взаємозв'язках дають більш точні оцінки. Для вирішення завдання оцінювання ймовірності страхових виплат в даній роботі пропонується використовувати логістичну регресію та байєсовський класифікатор. Для знаходження параметрів логістичної регресії пропонується використовувати різні методи залежно від виду вхідних даних.

Ключові слова: *особисте страхування, оцінка ймовірності, логістична регресія, байєсовський класифікатор, метод оцінки шансів, метод максимальної вірогідності.*

DETERMINING THE PROBABILITY OF INSURANCE POLICIES PAYMENTS BASED ON A DATA MINING METHODS

K.A. Bazilevich, M.S. Mazorchuk, A.A. Suhobrus

Socio-economic characteristics of policyholders have a great influence on the probability of insurance policies payments. The statistical data analysis techniques are the most commonly used for the determination of relevant factors and rate of influence. However, the recent wide application have had Data Mining methods, which provide a more accurate assessment in large amounts of information and complex relationships. In this paper it was proposed to use a logistic regression and Bayes classifierfor solving the problem of estimating the probability of insurance payments. It was proposed to use different methods for finding the parameters of the models depending on the type of input data.

Keywords: *private insurance, estimation of probability, logistic regression, Bayes classifier, method of assessingchances, the maximum likelihood method.*