Применение метода случайных лесов для оценки резерва произошедших, но еще не заявленных убытков страховой компании

Д.В. Денисов, Д.К. Смирнова

***Аннотация* - Целью настоящей работы является оценка применимости *метода случайных лесов* для оценки резерва произошедших, но еще не заявленных убытков (РПНУ) страховой компании по страхованию иному, чем страхование жизни. В основе задачи лежит статистический *метод случайных лесов* (Random forests). Для целей моделирования использовались реальные данные двух страховых компаний по прямому страхованию средств автотранспорта (КАСКО) за период 2009-2014 гг. Было проведено сравнение результата оценки РПНУ на 31.12.2014 г. *методом случайных лесов* с результатами расчетов стандартными методами (*цепной лестницы* и *Борнхютера-Фергюссона* по треугольникам оплаченных убытков). В целом, можно сделать вывод, что *метод случайных лесов* может быть применен для оценки РПНУ в качестве альтернативного алгоритма.**

***Ключевые слова* - моделирование убытков, РПНУ случайные леса, страхование.**

I. Введение

Задача страховой компании, которую необходимо решить в рамках имеющейся внутренней статистики, – это формирование страховых резервов, достаточных для исполнения обязательств по текущим договорам. По сути данная задача сводится к расчету резерва понесенных, но не заявленных убытков (РПНУ), поскольку расчет прочих резервов, как правило, либо производится общепринятыми методами, либо относится к уже известным, заявленным убыткам и тем самым не несет в себе фактора неопределенности. Расчет РПНУ требует наибольшей работы актуария в части анализа статистики и выбора метода оценки для адекватного прогнозирования будущих денежных потоков по уже произошедшим, но еще не заявленным в страховую компанию страховым случаям. Оценка РПНУ особенно важна для тех видов страхования, для которых характерно долгое урегулирование убытков.

Целью настоящей работы является оценка применимости *метода случайных лесов* для оценки РПНУ на основе реальных статистических данных по страхованию иному, чем страхование жизни.

Предметом исследования является *метод случайных лесов* (Random forests), относящийся к алгоритмам машинного обучения и предназначенный для решения задач классификации и регрессии, который был впервые представлен в статье [1]. Для применения *метода случайных лесов* мы будем использовать статистический пакет R, при помощи функций библиотеки randomForest.

II. Метод случайных лесов

A. Теоретический обзор

С помощью бутстрепа (bootstrap) - выборки с возвращением, - на основе каждого из случайно выбранных подмножеств тренировочной выборки строится свое дерево принятия решений. Параметрами метода являются:

1) Количество деревьев принятия решений в ансамбле - N.
2) Число случайно отбираемых признаков обучающей выборки для построения деревьев – m из M признаков исходного множества.

Оценка регрессии получается в результате усреднения оценок регрессии всех деревьев. Пусть решения, принимаемые каждым отдельным деревом, будут не самыми лучшими, однако «лес» деревьев может принимать вполне разумные решения.

При построении каждого отдельного дерева при определении тренировочной и тестовой выборки используется бэггинг (bagging): выборка случайных двух третей наблюдений в качестве обучающей выборки и одна треть для оценки результата.

Метод случайных лесов можно описать следующим образом [2].

1) Для каждого из N деревьев в ансамбле ($j = 1, 2, \ldots N$):
   - Формируется бутстреп-выборка S размера k по исходной обучающей выборке $\Omega = \{x_i, y_i\}_{i=1}^{k}$.
   - По бутстреп-выборке S к неусечённому дереву решений T_j рекурсивно применяются следующие шаги:
     (a) Случайным образом выбираются m из M имеющихся объясняющих переменных.
     (b) Из отобранных m переменных выбирается признак, наилучшим образом обеспечивающий расщепление вершины согласно классическому алгоритму CART. Вершина расщепляется согласно данному признаку на две подвыборки.

В результате получаем лес деревьев решений $\{T_j\}_{j=1}^{N}$.

Статья получена 27 мая 2016. Работа представляет собой результат магистерской диссертации.
Д.В. Денисов, к.ф.-м.н., МГУ им. М.В. Ломоносова.
Д.К. Смирнова, магистр, МГУ им. М.В. Ломоносова.

2) Значения зависимой переменной для новых наблюдений предсказываются согласно следующей формуле:

$$\hat{f}_{rf}^{N} = \frac{1}{N}\sum_{j=1}^{N} T_j(x) \quad (1)$$

От классического метода построения деревьев решений *метод случайных лесов* отличает установка количества отбираемых признаков и то, что каждое дерево леса строится без усечений.

B. Метод «Out-Of-Bag» оценки ошибки прогноза

Оценка ошибки прогноза случайного леса осуществляется методом «Out-Of-Bag» (OOB) [4]. При использовании бутстрепа примерно 37% наблюдений исходной обучающей выборки не используются для построения деревьев решений (так как выборка с возвращением не содержит некоторые наблюдения, а некоторые наоборот попадают в нее несколько раз). Для целей регрессии некоторого вектора x используются только те деревья леса, которые строились по бутстреп выборкам, не включающим в себя оцениваемый вектор.

C. Оценка ошибки прогноза модели случайного леса

Имеется уравнение регрессии следующего вида:

$$y \sim X\beta, \quad (2)$$

где n – объем выборки,
k – количество объясняющих переменных в модели, причем $k \ll n$,
y – вектор значений зависимой переменной (оплаченных убытков), размерности $(n \times 1)$,
β – вектор коэффициентов модели, размерности $(k \times 1)$,
X – матрица объясняющих переменных, размерности $(n \times k)$,

Для оценки качества ***модели случайного леса*** для целей регрессии в библиотеке randomForest в среде R используются следующие критерии [5].

MSE (Mean square errors) – вектор среднеквадратических ошибок длины n, где n – количество наблюдений. Для каждого критерия суммирование ведется по всем $i = \overline{1,n}$. Для задачи регрессии величина MSE, вычисленный при помощи метода «Out-Of-Bag», является оценкой ошибки прогноза модели:

$$MSE^{OOB}(y) = \frac{1}{n}\sum_{i}(y_i - \hat{y}_i^{OOB})^2, \quad (3)$$

где y_i – значения зависимой переменной исходной выборки,
$\hat{y}_i^{OOB}$ – среднее из предсказанных методом OOB значений наблюдений.

% Var explained (Percent variance explained) – процент объясненной дисперсии:

$$\%Var(y) = \left(1 - \frac{\sum_i (y_i - \hat{y}_i^{OOB})^2}{\sum_i (y_i - \bar{y})^2}\right) * 100\% \quad (4)$$

где $\bar{y}$ – среднее значение зависимой переменной.

Чем ниже значение MSE^{OOB} и $\%Var$, тем выше качество модели.

D. Меры информативности

Для целей регрессии в библиотеке randomForest программы R используются две меры информативности (important measures), которые помогают выделить наиболее информативные признаки модели для задачи классификации.

Пусть x_i – рассматриваемая переменная. Информативность x_i оценивается исходя из того, как меняется ошибка прогноза при изменении значений данной переменной при неизменности значений остальных переменных. Необходимые расчеты производятся для каждого дерева по ходу построения леса [6].

Мера 1. Вычисления величины меры 1 можно представить в виде выполнения четырех шагов:

1. Построение случайного леса и получение ошибки прогноза MSE^{OOB};
2. Модификация OOB выборок путем перестановки значений признака x_i для каждого дерева из леса;
3. Вычисление оценки ошибки прогноза $\widetilde{MSE}^{OOB}$ по модифицированным выборкам;
4. Информативность признака x_i определяется по формуле:

$$\%\text{IncMSE}(x_i) = \frac{1}{N}\frac{\sum_{j=1}^{N}(\widetilde{MSE}_i^{OOB} - MSE_i^{OOB})}{\widehat{Var}(MSE^{OOB})} = \frac{\sum_{j=1}^{N}(\widetilde{MSE}_i^{OOB} -$$
$$\sum_{j=1}^{N}(MSE_i^{OOB} -$$

где $\%\text{IncMSE}(x_i)$ – значение меры 1 для i-го признака.

Мера 2. Второй мерой является суммарное уменьшение критерия MSE во всех вершинах деревьев леса вследствие расщепления вершины на основе данной переменной, усредненное по всем деревьям ансамбля.

$$IncNodePurity(x_i) =$$
$$= \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{t_j}(MSE_{j,k}^{OOB} - MSE_{i,j,k}^{OOB\ after\ splitting}), \quad (7)$$

где $IncNodePurity(x_i)$ – значение меры 2 для i-го признака.
N – количество деревьев в ансамбле;
t_j – количество вершин в j-м дереве, $j = 1 \ldots N$;
$MSE_{j,k}^{OOB}$ – критерий MSE^{OOB} для k-й вершины j-го дерева до расщепления;
$MSE_{i,j,k}^{OOB\ after\ splitting}$ – критерий MSE^{OOB} для j-го дерева после расщепления k-й вершины на основании i-го признака.

Для обеих представленных мер информативности верно: чем выше значение меры, тем выше значимость рассматриваемого признака для данной модели. Тем не менее, как правило, переменным, имеющим большее количество уникальных значений, чаще соответствуют более высокие значения меры [7].

III. Модель оценки оплаты заявленного убытка

A. Описание данных

Для целей моделирования будем использовать реальные данные двух страховых компаний (далее – СК №1 и СК №2) по виду страхования «Страхование средств автотранспорта» (далее – «КАСКО», «резервная группа»). В целях конфиденциальности наименования компаний не раскрывается. Имеется статистика для периода 2009-2014 гг. по СК №1 и для периода 2010-2014 гг. по СК №2.

Далее в тексте употребляются следующие обозначения:

Категориальные переменные.

region филиал, к которому относится полис
start_quar квартал начала действия полиса

Числовые переменные.

end_year год окончания действия полиса
term_end срок действия договора, в днях
claim_date дата поступления заявления
claim_year год поступления заявления
ins_sum размер страховой суммы, руб.
paid величина оплаты заявленного убытка, руб.
claim_delay задержка в поступлении заявления об убытке, исчисленная от момента начала действия договора, в днях:

$$claim\_delay == \begin{cases} claim\_date - start\_date \text{ если } claim > 0; \\ term\_end + 1; \text{ если } claim = 0. \end{cases}$$

paid_edited величина оплаты, если убыток был заявлен в год, следующий после года начала действия договора - числовая переменная:

$$paid\_edited = = \begin{cases} paid: \text{ если } claim\_year > start\_year: \\ 0; \text{ иначе.} \end{cases}$$

year_of_ins_ev год наступления страхового случая

Фиктивные переменные.

crisis_year_of_ ins_ev фиктивная переменная, обозначающая кризисный год наступления страхового случая

$$crisis\_year\_of\_ins\_ev == \begin{cases} 1; \text{ если } year\_of\_ins\_ev = \{2009\}; \\ 0; \text{ иначе.} \end{cases}$$

B. Предпосылки модели

Всего за исключением досрочных расторжений имеется 3584 наблюдений по СК №1 и 1380 наблюдений по СК №2. Структура СК №1 представляет собой центральный офис и 13 филиалов, в то время как СК №2 не имеет филиалов и представительств. Наблюдения за весь период наблюдений будем рассматривать как стационарные данные.

Для моделирования размера оплаты заявленного убытка мы делаем следующие предположения:

1) По каждому договору имеет место один убыток. Если по договору по факту произошел более, чем один убыток, мы суммируем данные по колонке «Величина оплаты». При этом заявленный убыток и страховой случай соответственно датируются наиболее ранней датой из представленных дат.
2) Убыток не может быть заявлен по расторгнутому договору.
3) Практически все страховые договоры по КАСКО действуют ровно один год или менее. Поэтому можно с достаточной степенью уверенности утверждать, что для рассматриваемой резервной группы данные об убытках по договорам, дата окончания срока действия которых не позднее отчетного года, будут достоверным основанием для прогнозирования будущих выплат.
4) Построение и тестирование модели производилось на основе тех наблюдений, для которых дата окончания действия полиса в силу не позднее 2014 г., так как для таких договоров мы знаем статистику наступления страховых случаев в год, следующий за годом вступления договора в силу. Данные же договоров, дата окончания действия которых позднее 2014 г., использовались для прогнозирования.
5) В силу того, что доля досрочных расторжений договоров составляет менее 1%, расторжениями в рассматриваемой модели можно пренебречь.

C. Постановка задачи оптимизации

В настоящей работе ставится следующая задача оптимизации:

$$\widehat{paid\_edited}_{i_{end\_year}} - paid\_edited_{i_{end\_year}} \quad (11)$$
$$для\ end\_year = 1, \dots R$$

Где R – номер отчетного года (периода),
i_{end_year} – номер наблюдения (из имеющейся статистики) для года *end_year*,
$paid\_edited_{i_{end\_year}}$ – реальная величина оплаты заявленного убытка для наблюдения i_{end_year},
$\widehat{paid\_edited}_{i_{end\_year}}$ – оцененная величина оплаты заявленного убытка для наблюдения i_{end_year}, причем $paid\_edited_{i_{end\_year}} \geq 0$ и $\widehat{paid\_edited}_{i_{end\_year}} \geq 0$.

Целью настоящей работы является прогноз величины оплаты заявленного убытка для отчетного года $\widehat{paid\_edited}_{i_R}$ для таких наблюдений i_R, для которых выполнено:

$$i_R \in \{1, \dots n_R\}: \begin{cases} end\_year_{i_R} > R \\ paid_{i_R} = 0 \end{cases} \quad (12)$$

Где n_R – количество наблюдений для года R.
При этом оцененная величина $\widehat{paid\_edited}_{i_R}$ для отчетного года R будет представлять собой прогноз предстоящих в следующем после отчетного года страховых выплат по договорам, заключенным не позднее отчетного года R. Таким образом, величина $\widehat{paid\_edited}_{i_R}$ будет представлять собой сумму всех страховых резервов компании. Для получения величины РПНУ необходимо будет вычесть прочие страховые резервы из расчетного значения $\widehat{paid\_edited}_{i_R}$.

D. Модель оценки размера оплаты методом случайных лесов

Подробное построение модели оценки будущих выплат представлено на примере ***СК №1.*** Разобьем выборку на две составляющие: тестовую и обучающую. В качестве обучающей выборки случайным образом выберем 70% от количества наблюдений исходного множества, зафиксировав при этом зерно случайного процесса, то есть $0{,}7 * 3507 = 2455$, для тестовой – аналогично выберем случайные $3507 - 2455 = 1052$ наблюдений.

Будем оценивать следующую зависимость:

$$paid\_edited \sim crisis\_year\_of\_ins\_ev + ins\_sum + term\_end + start\_q$$

Для определения оптимального количества деревьев в ансамбле построим модель с использованием 2000 деревьев, выводя с шагом 100 деревьев значения оценки ошибки прогноза. Количество деревьев с наименьшим значением ООВ-ошибки MSE^{OOB} будет искомым оптимальным значением параметра *N*.

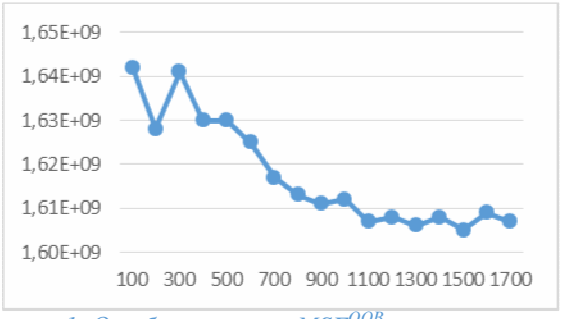

Рисунок 1. Ошибка прогноза MSE^{OOB} в зависимости от количества деревьев в ансамбле.

Как видно из графика, наименьшая ошибка прогноза наблюдается при количестве деревьев в ансамбле $N = 1500$.

При помощи функции tuneRF определим оптимальное количество переменных для расщепления вершин [8] при значении параметра «количество деревьев в ансамбле» $N = 1500$. Как видно из таблицы ниже, наименьшая ошибка прогноза наблюдается при расщеплении вершин на основе четырех случайно выбранных переменных: $m = 4$.

Таблица 1. Прогностическая точность модели в зависимости от числа переменных для разбиения

	MSE^{OOB}
1	1 767 193 652
2	1 629 644 628
4	**1 610 194 250**

Исходя из рекомендаций по установке параметров, построим ***Модель 1:*** $N = 1500, m = 4$.

При переобучении характерно, что ошибка прогноза на тестовой выборке дает большее значение, чем на обучающей. Воспользуемся оценкой MSE^{OOB} и $\%Var$ для ***Модели 1***, используя метод "Out-of-bag". Для тестовой выборки ***Модели 1*** $MSE^{OOB} = 1{,}4 * 10^9$ и $\%Var = 5{,}31\%$, в то время как для обучающей $MSE^{OOB} = 1{,}6 * 10^9$ и $\%Var = 22{,}58\%$. Можем убедиться, что полученная модель не переобучена.

Оценим значения мер информативности для переменных, участвующих для обучающей выборки ***Модели 1***.

Чем выше значение Меры 1 и Меры 2 для объясняющей переменной, тем больше она влияет на оценку регрессии объясняемой переменной. Как видно на рисунке выше, для различных переменных Мера 1 и Мера 2 определяют информативность по-разному. Так как Мера 1 не принимает отрицательных значений, можем убедиться, что все независимые переменные являются значимыми и их исключение не улучшит модель.

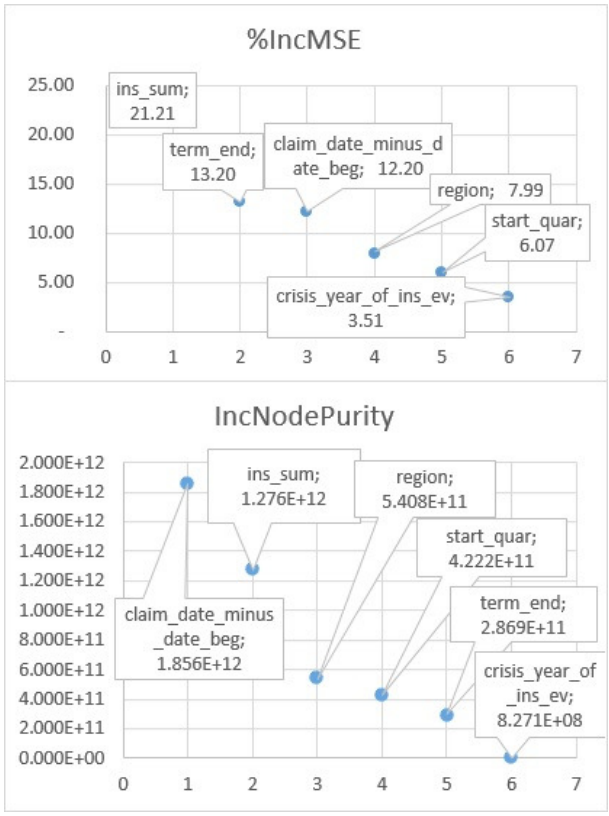

Рисунок 2. Информативность независимых переменных Модели 1 (N=1500, m=4)

Мы получили прогноз величины оплаты заявленного убытка. Ниже представлено краткое описание полученной модели:

Таблица 2. Описание Модели 1 для Страховой компании №1

Базовые предположения модели	$N = 1500, m = 4$; объясняющие переменные: *crisis_year_of_ins_ev, ins_sum, term_end, start_quar, region, claim_delay.*	$N = 1500, m = 4$; объясняющие переменные: *crisis_year_of_ins_ev, ins_sum, term_end, start_quar, region, claim_delay.*
Специальные предположения модели	*crisis_year_of_ins_ev*, (т.е. предполагается, что 2014 г. – не кризисный год)	*crisis_year_of_ins_ev*, (т.е. предполагается, что 2014 г. – кризисный год)
Прогноз будущих выплат методом случайных лесов	1 471 280 руб.	1 420 415 руб.

Для получения величины РПНУ, из полученного прогноза выплат мы должны вычесть все прочие страховые резервы. Суммы прочих страховых резервов приведены в соответствии с актуарным заключением рассматриваемой СК за 2014 г.

Таблица 3. Расчет РПНУ, исходя из оценки прогноза будущих выплат методом случайных лесов для СК №1

Специальные предположения модели / **Показатель**	*crisis_year_of_ins_ev =FALSE*, (т.е. предполагается, что 2014 г. – не кризисный год)	*crisis_year_of_ins_ev =TRUE*, (т.е. предполагается, что 2014 г. – кризисный год)
Прогноз будущих выплат методом случайных лесов, руб. (1)	1 471 280	1 420 415
Резерв незаработанной премии, руб. (2)	802 932	802 932
Резерв неистекшего риска, руб. (3)	25 990	25 990
Резерв заявленных, но не урегулированных убытков, руб. (4)	129 126	129 126
РПНУ, руб. (5) = (1) - (2) - (3) - (4)	**513 232**	**462 367**

Тот факт, что прогноз будущих выплат и, в частности, прогноз РПНУ по КАСКО при условии, что 2014 г. – является кризисным годом, по сумме меньше, чем если считать 2014 г. не кризисным, является логичным, так как в условиях кризиса население склонно отказываться от личного автотранспорта в пользу общественного.

Аналогичным образом был проведен подбор параметров и построение модели *методом случайных лесов* для ***СК №2.*** Так как в структуре страховой компании №2 отсутствуют филиалы и представительства, в модели отсутствует объясняющая переменная *region.* Также в модели отсутствует объясняющая переменная *crisis_year_of_ins_ev,* так как почти все договоры из имеющейся статистики представлены, начиная с 2010 г. Ниже представлено краткое описание полученной модели.

Таблица 4. Описание Модели 2 для СК №2

Базовые предположения модели	$N = 1600, m = 1$; объясняющие переменные: *ins_sum, term_end, start_quar, claim_delay.*
Специальные предположения модели	отсутствуют
Прогноз будущих выплат методом случайных лесов	4 469 469 руб.

Для получения величины РПНУ, из полученного прогноза выплат мы должны вычесть все прочие страховые резервы. Суммы прочих страховых резервов приведены в соответствии с актуарным заключением рассматриваемой СК за 2014 г.

Таблица 5. Расчет РПНУ, исходя из оценки прогноза будущих выплат методом случайных лесов для СК №2

Прогноз будущих выплат методом случайных лесов, руб. (1)	4 469 469
Резерв незаработанной премии, руб. (2)	3 568 623
Резерв неистекшего риска, руб. (3)	0
Резерв заявленных, но не урегулированных убытков, руб. (4)	0
РПНУ, руб. (5) = (1) - (2) - (3) - (4)	900 846

IV. Сравнение результатов оценки РПНУ различными методами

При расчете РПНУ *методом Борнхютера-Фергюссона*. (далее – «БФ метод») для ***СК №1*** наблюдается низкий, а для ***СК №2*** – наоборот высокий уровень выплат в последних кварталах, а также в обеих компаниях наблюдается нестабильность коэффициентов *метода факторов развития*. В связи с этим коэффициент убыточности *k*, предполагаемый для последних кварталов, был выбран на уровне отношения:

$$\frac{\text{Резерв убытков} + \text{Конечная величина убытка}}{\text{Заработанная премия}}, \quad (14)$$

При этом для ***СК №1*** коэффициент убыточности *k* для всех четырех кварталов 2014 г. был выбран в размере 110% на уровне отношения (14), где все величины брались за период 2013-2014 гг. Для ***СК №2*** коэффициент убыточности *k* только для 4го квартала 2014 г. был выбран в размере 90% на уровне отношения (14), где все величины брались за период 2012-2013 гг.

Ниже представлено сравнение результатов расчета РПНУ различными методами с реальными данными рассматриваемых СК за 2015 г.

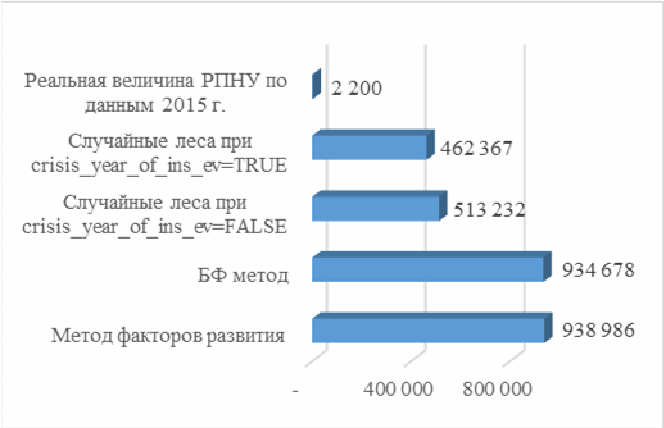

Рисунок 3. Сравнение результатов расчета РПНУ с реальными данными СК №1

Следует отметить, что *метод факторов развития* не работает, если статистика происшедших и оплаченных убытков ненадежна, например, в случае, когда заработанная премия нестабильна, и объем портфеля заключенных договоров меняется от года к году. БФ метод дает возможность скорректировать вручную размер РПНУ, исходя из изменений заработанной премии в последние годы, при этом актуарий руководствуется, как правило, своим

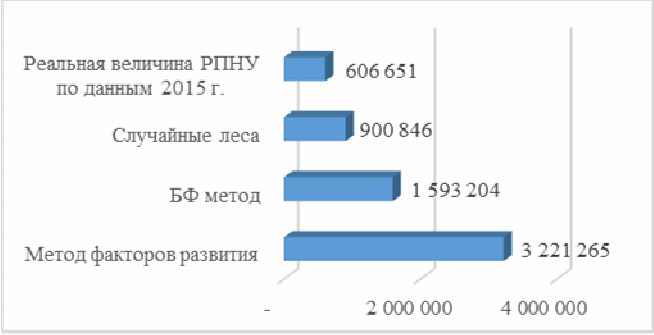

Рисунок 4. Сравнение результатов расчета РПНУ с реальными данными СК №2

профессиональным мнением при выборе коэффициентов.

Метод случайных лесов дает результаты более близкие к реальным данным, чем стандартные методы расчетов, однако все равно завышенные. Следует отметить, что более консервативная оценка страховых резервов является предпочтительней для надзорных органов, особенно в условиях нестабильной экономической конъюнктуры. Можно сделать вывод, что на указанных данных *метод случайных лесов* прогнозирует будущие выплаты лучшим образом, чем стандартные методы.

Важно отметить, что моделирование *методом случайных лесов*, в том виде, в котором оно представлено в настоящей работе, применимо для оценки будущих страховых выплат только к тем резервным группам, для которых характерно долгое урегулирование убытков, так как представленные модели базируются на статистике оплат, произведенных в году, следующем за годом вступления полиса в силу. Кроме того, данный метод аналогично *методу цепной лестницы* и *БФ методу* требует большого количества статистики для прогнозирования.

V. Заключение

В настоящей статье мы применили *метод случайных лесов* для оценки резерва происшедших, но не заявленных убытков страховой компании для резервной группы «Страхование средств автотранспорта». Исследование показало, что оценка *методом случайных лесов* дает завышенные результаты по сравнению с реальными данными, однако дает прогноз более близкий к реальным данным, чем оценка стандартными методами. Таким образом, сумма будущих выплат, оцененная *методом случайных лесов*, может быть использована актуарием в качестве проверки адекватности суммы всех страховых резервов, рассчитанных стандартными методами.

Основным преимуществом *метода случайных лесов* является относительная легкость настройки вводных параметров: количества деревьев в ансамбле и количества переменных, используемых для расщепления вершин деревьев. При этом выбор параметров модели стандартизирован и не зависит от профессионального суждения актуария. Модель *случайных лесов* способна выявлять сложные нелинейные взаимосвязи между переменными. Кроме того, *метод случайных лесов* имеет встроенный алгоритм оценки ошибки прогноза (на основе OOB выборок). Среди других важных достоинств метода следует отметить устойчивость к «выбросам», отсутствие необходимости нормировать или иным образом преобразовывать данные (поддерживается работа с категориальными переменными), возможность реализовать данный алгоритм на основе параллельных вычислений, что важно при больших объемах данных.

Модель, построенная *методом случайных лесов*, является более гибкой, так как дает возможность актуарию включить различные параметры, не учтенные в стандартных методах, причем в разных комбинациях.

Тем не менее, *метод случайных лесов* обладает и рядом недостатков: сложность интерпретации результатов (так как невозможно точно понять, насколько и через какие параметры повлияло то или иное неоптимальное дерево на результат «голосования»), невозможность визуализации решения, склонность к «переобучению» на некоторых задачах (особенно на зашумленных). Модель чаще всего получается громоздкой, так как содержит большое количество построенных деревьев. *Метод случайных лесов* аналогично стандартным методам требует для прогнозирования достаточно большой статистики по убыткам.

В целом можно сказать, что *метод случайных лесов* может быть применен для оценки резервов страховых компаний по страхованию иному, чем страхование жизни в качестве альтернативного алгоритма.

С помощью применения *метода случайных лесов* в совокупности со стандартными методами, актуарий может представить интервал оценок страховых резервов, в рамках которого руководство страховой компании может принимать ответственные финансовые решения. Таким образом, более тонкая «настройка» такого финансово-экономического инструмента, как страховые резервы, при грамотном использовании помогает разрешить основной конфликт страхового бизнеса: между обеспеченностью обязательств и конкурентоспособностью на волатильном рынке.

Библиография

[1] Breiman L. *Random forests* // Machine learning. 2001. *45*(1). P. 5-32.
[2] Siroky D. *Navigating Random Forests and related advances in algorithmic modeling* // Statistics Surveys, 3. 2009. P. 147–163.
[3] Breiman L., Friedman R., Olshen R., & Stone C. *Classification and Regression Trees* // Belmont, California: Wadsworth International. 1984.
[4] Breiman L. *Out-of-bag estimation* // Berkeley: Technical Report, Statistics Department University of California. 1996. 13 p.
[5] Liaw A., Wiener M. *Classification and Regression by Random Forest* // R News, 2 (3). 2002. P. 18-22.
[6] Чистяков С. *Случайные леса: обзор* // Труды Карельского научного центра РАН (1). 2013. С. 117-136.
[7] Груздев А.В. *Метод случайного леса в скоринге* // Риск-менеджмент в кредитной организации (№1 (13)). 2014. С. 28-43.
[8] Breiman L. *Manual on setting up, using, and understanding random forests v 4.0.* URL (дата обращения 22.10.2015 г.): https://www.stat.berkeley.edu/~breiman/papers.html

Application of Random forest method to estimate the incurred but not reported claims reserve of an insurance company

D.V. Denisov, D.K. Smirnova

Abstract **- The purpose of this report is to explore the applicability of *Random forest method* to assess the incurred but not reported claims reserve (IBNR) of a non-life insurance company. The research is based on the statistical method of Random forest. The actual data on the direct hull insurance of two real companies for the period 2009-2014 were used. The IBNR valuated on 31.12.2014 by *Random forest* was compared with the results of standard calculation methods (*chain ladder* and *Bornhuetter – Ferguson* on paid triangles). In general, we can say that the *Random forest method* can be applied to assess the IBNR as an alternative algorithm.**

Key words **– loss modelling, IBNR, random forests, insurance.**