ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

**Факультет экономических наук**

**Образовательная программа "** **Прикладная экономика"**

##### **Эльтекова Дарья Дмитриевна**

***«Подходы к оценке вероятности дефолта на примере сегмента страховых компаний»***

##### Выпускная квалификационная работа - МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

по направлению подготовки 38.04.01 Экономика

|  |  |
| --- | --- |
| **Рецензент**  **Малахов**  **Дмитрий**  **Игоревич** | **Руководитель**  **Дробин**  **Сергей**  **Алексеевич** |
|  |  |

**Москва 2019**

**Оглавление**

[**1. Введение** 4](#_Toc9892461)

[**2. Обзор литературы** 8](#_Toc9892462)

[**3. Описание рынка страхования в России за период с 2011 по 2018 года** 12](#_Toc9892463)

[**4. Данные для эмпирического исследования** 16](#_Toc9892464)

[**4.1. Описание логит-регрессии** 20](#_Toc9892465)

[**4.2. Метрика оценивания модели** 22](#_Toc9892466)

[**4.3. Анализ финансовых переменных** 23](#_Toc9892467)

[**4.3.1. Показатели финансовой устойчивости (financial stability)** 24](#_Toc9892468)

[**4.3.2. Показатели Рентабельности (profitability)** 27](#_Toc9892469)

[**4.3.3 Показатели убыточности (unprofitability)** 30](#_Toc9892470)

[**4.3.4. Показатели оценки перестраховочных операций и инвестиционной деятельности (reinsurance operations and investment activity)** 31](#_Toc9892471)

[**4.3.5 Показатели оценки платежеспособности страховой компании и оценки ее ликвидности (liquidity and solvency)** 33](#_Toc9892472)

[**4.3.6. Показатели страховой деятельности и доли рынка** 36](#_Toc9892473)

[**4.3.7. Обработка данных** 41](#_Toc9892474)

[**4.3.8. Отбор факторов** 42](#_Toc9892475)

[**4.4. Дамми-показатели** 50](#_Toc9892476)

[**5. Моделирование** 53](#_Toc9892477)

[**5.1. Логит-модель от всех переменных (Модель 0)** 53](#_Toc9892478)

[**5.2. Логит-модель от всех значимых переменных (Модель 1)** 55](#_Toc9892479)

[**5.3. RFECV метод для Логит-модели (Модель 2.1 и Модель 2.2)** 57](#_Toc9892480)

[**5.4. Перебор логит-моделей (Модели 3.1, 3.2 и 3.3)** 61](#_Toc9892481)

[**5.5. Модель с машинным обучением (Модель 4)** 66](#_Toc9892482)

[**5.6. Сравнение моделей** 68](#_Toc9892483)

[**5.7. Описание итоговой модели** 71](#_Toc9892484)

[**6. Прогноз на 2019 год** 74](#_Toc9892485)

[**7. Заключение** 75](#_Toc9892486)

[**8. Литература** 77](#_Toc9892487)

[**9. Приложение** 81](#_Toc9892488)

# **1. Введение**

В магистерской работе были исследованы различные модели, предсказывающие вероятность дефолта российских страховых компаний, с использованием логистической модели бинарного выбора и дерева решений. Рассмотрен обширный ряд моделей вычисления вероятности дефолта для страховщиков применительно к российской действительности на основе годовой финансовой отчетности и сводных отчетных данных по страховой деятельности за период с 2012 по 2016 гг. Так же к объектам исследований добавлены бинарные факторы вида страховой деятельности, осуществляемой за отчетный период, региональный фактор и год. Кроме этого произведена валидация модели на данных 2017 года и дефолтах за весь 2018 г. Актуальность работы заключается в том, что вышеперечисленные модели помогут предсказывать дефолты российских страховых компаний и, следовательно, окажутся востребованными как для страхового сектора и регулятора, так и для потребителей страховых услуг.

Для достижения поставленной цели в работе были поставлены и решены следующие задачи:

• исследовать существующие теоретические и практические подходы по оценке вероятности дефолта страховых компаний;

• проанализировать бухгалтерский баланс и отчет о прибыли и убытках ряда страховых компаний, виды предоставляемых ими услуг за несколько лет, и изучить общие тенденции для всех страховых компаний;

• сделать обзор и выбрать ключевые показатели и коэффициенты деятельности страховщиков, которые лучше всего классифицируют компании;

• вычислить относительные и абсолютные факторы для ряда компаний за несколько лет и собрать данные в сводную таблицу;

• провести однофакторный анализ ключевых показателей финансовой и страховой деятельности;

• разделить выборку на обучающую и тестовую. По данным за прошлые года отмечая, какие компании пришли в убыток, или у них была отозвана лицензия, а какие наоборот, из года в год успешно ведут свою деятельность, провести эконометрический анализ, сравнить и выбрать наилучшую модель, провести валидацию модели на тестовых данных.

• Обучить модель на всех данных и по данным годовой финансовой отчетности за 2018 год построить прогноз на 2019 год.

В современном мире особо важную роль играют страховые организации, компенсирующие разного рода убытки и ущербы как физических, так и юридических лиц за счет взносов в страховые фонды. Поскольку возможный ущерб носит вероятностный характер, при чем число пострадавших меньше всех участников страхования, то происходит перераспределение страхового фонда как в пространстве, так и во времени. Для минимизации рисков потерь при наступлении различных неблагоприятных событий хозяйствующий субъект может за небольшую плату воспользоваться услугами страхования и иметь уверенность, что в случае наступления страхового случая страховщик компенсирует весь ущерб.

Сектор страхования играет важную роль в финансовом секторе. Во-первых, страховое покрытие позволяет организациям, корпорациям и предприятиям снимать с себя риски. Например, страховые компании помогают фирмам и домашним хозяйствам ограничивать финансовые издержки, связанные с возникновением различных рисков, касающихся их физического имущества, юридической ответственности и различных финансовых потерь. Во-вторых, страховые компании направляют сбережения в инвестиции. Например, страховые компании, занимающиеся страхованием жизни, помогают людям покрывать риски, связанные с неопределенностью в отношении их здоровья и продолжительности жизни, и одним из способов их осуществления является сбор средств со стороны страхователей и их инвестирование в долг, капитал и другие активы.

Для страховой деятельности характерен существенный временной разрыв между приемом страхового взноса и временем предоставления страховой услуги, а именно страховых выплат. Поэтому хозяйствующее лицо должно быть уверенно, что по истечении времени у страховой организации не возникнет трудностей в положенное время выплатить всю компенсацию в полном объеме, как это было зафиксировано в договоре.

Устойчивость служит залогом выживаемости и продолжительного стабильного существования организации на рынке, а также является отражением стабильного превышения доходов над расходами. Если страховая компания находится в неустойчивом положении, то отрицательное влияние внешних и внутренних факторов может привести к ухудшению ее финансового состояния. Для анализа текущего, инвестиционного и финансового развития организации проводится анализ его финансовой устойчивости, который отражает способность предприятия отвечать по своим долгам и обязательствам.

Страховые компании должны вести свою деятельность не прекращая так, чтобы это не привело к их разорению, при этом строго соблюдая свои обязательства перед страхователями, чем обеспечивая надёжную страховую защиту хозяйствующих субъектов. Это регулируется на юридическом уровне, и минимальные требования к страховым организациям закреплены законодательно. Так в статье 25 закона РФ от 27.11.1992 N 4015-1 (ред. от 31.12.2017) "Об организации страхового дела в Российской Федерации" перечислены условия обеспечения финансовой устойчивости и платежеспособности страховщика, такие как: “экономически обоснованные страховые тарифы; сформированные страховые резервы; средства страховых резервов, достаточные для исполнения обязательств по страхованию, сострахованию, перестрахованию, взаимному страхованию; собственные средства (капитал) и перестрахование”. [1] Минимальный размер уставного капитала страховщика фиксируется на уровне 120 млн. рублей при учете коэффициентов, равных одному, двум или четырем, в зависимости от вида деятельности. Так же по указанию Банка России установлен (в п. 4.1 Федерального закона от 23.07.2013 N 234-ФЗ (ред. от 09.01.2018)) порядок расчета страховой организацией нормативного соотношения собственных средств (капитала) и принятых обязательств. Он рассчитывается страховой организацией как отношение фактического размера маржи платежеспособности к нормативному размеру и отражает достаточность собственных средств (капитала) страховой организации. [2] Методика расчета фактического и нормативного размера платежеспособности для разных видов страхования приведена и в Приказе Минфина РФ от 02.11.2001 N 90н. [3] Для упрощения надзора за страховщиками раздел со сведениями о выполнении требований финансовой устойчивости и платежеспособности обязан содержаться в аудиторском заключении о годовой бухгалтерской (финансовой) отчетности страховщика (ст. 29 Закона РФ от 27.11.1992 N 4015-1). [1]

# **2. Обзор литературы**

Принимая во внимание тот факт, что прогнозирование банкротства является важной и широко изучаемой темой, мы сосредоточим наш обзор литературы на прогнозировании банкротства страховой компании с использованием финансовых показателей. Кроме того, мы сосредоточимся на зарубежных и отечественных исследованиях, в которых реализован logit-model подход для оценивания вероятности наступления дефолта.

История прогнозирования банкротства произошла от предсказания банкротства предприятий. Важный вклад Altman (1968) побудил исследователей использовать многомерный анализ для прогнозирования банкротства фирм. [4] Он представил оригинальную формулу Z-показателя (1968 г.) и продемонстрировал ее преимущество, проанализировав пять основных финансовых и экономических аспектов деятельности фирмы: показатели ликвидности, размера; операционная эффективность и прибыльность активов, финансовый рычаг, а также учет возможностей управления в условиях конкуренции (общий оборот активов). Синки (1975) использовал дискриминантный анализ для прогнозирования банкротств банков. [5]

Для сравнения, Martin (1977) и Ohlson (1980) использовали логистическую регрессию для прогнозирования банкротств фирм и банков. [6] [7] Мартин (1977) попытался предсказать банкротство коммерческого банка США в течение 2 лет в течение 1970 и 1976 годов, используя 25 финансовых коэффициентов риска активов, ликвидности, достаточности капитала и прибыли. Он предположил, что логистическая регрессия имеет более высокий процент правильно классифицированных, чем линейный дискриминант. После этих первоначальных исследований были проведены эмпирические исследования для сравнения точности прогнозирования этих двух подходов (Boyacioglu et al., 2009). [8]

Так в статьях Карминского-Кострова (2012, 2013, 2017) рассмотрены различные logit-модели вероятности дефолта российских банков. [9, 10, 11] Для моделей были взяты относительные финансовые переменные из квартальных финансовых отчетов. Все переменные с лагом в два квартала были разделены на 5 групп, после чего было обозначено их влияние на вероятность дефолта. В статье авторы выдвинули гипотезы о том, что переменные, представленные отношением капитала или банковской прибыли к активам и размером активов входят в модель нелинейно. В результате исследования и сравнения различных моделей авторы отмечают, что при добавлении в модель квадратов финансовых переменных, имеющих U-образную зависимость с вероятностью дефолта, дамми-переменной года, макро- и институциональных показателей качество модели заметно улучшается. А оценивание модели с использованием logit-модели панельных данных к улучшению предсказательной силы не привело.

В исследовании (BarNiv, McDonald, 1992) авторы представляют методологический подход для выявления несостоятельных страховых компаний в США в период с 1961 по 1988 гг. [12] В этой статье финансовые бедствия и несостоятельность используются взаимозаменяемо для описания страховщиков, переживающих ликвидацию. Авторы представили достаточно надежную logit- модель для классификации и прогнозирования финансовых затруднений в страховой отрасли, а также предпринята попытка решения методологических вопросов, которых предыдущие исследования иногда игнорировали. После отбора коэффициентов по их классифицирующей силе авторы из 45 объясняющих переменных оставили 7, такие как: чистая прибыль к общей сумме активов, размер капитала, чистая прибыль к капиталу и размер обязательств. Предполагается, что эти переменные связаны со страховщиками, которые могут стать неплатежеспособными.

В статье (Chen, Wong, 2004) основное внимание уделяется состоянию здоровья компаний по страхованию в Азии с использованием данных фирмы и макроданных по отдельности в разрезе разных стран. [13] Факторы, которые являются существенными для оценки неплатежеспособности страховщиков, включают размер фирмы, эффективность инвестиций, ликвидность, операционную маржу, рост премий и сальдо роста. Авторы в своем исследовании построили Logit – модель, которая показывает вероятность банкротства компании с помощью бинарной зависимой переменной, которая является функцией вектора объясняющих переменных. Модель для каждой страны дает различную комбинацию значимых предикторов финансового здоровья общих страховщиков. Это может быть во многом обусловлено различием в экономической, социальной, юридической и политической обстановке, что приводит к различным детерминантам финансовой устойчивости основных страховщиков с разной значимостью. Таким образом, регулирующие органы в разных странах могут давать разные веса при измерении финансового состояния своих общих страховщиков. Более того, различные временные эффекты могут также повлиять на прогноз финансовой устойчивости страховщиков. Этот интересный вывод из международного исследования может относится и ко всем предыдущим работам.

В статье (Le, Viviani, 2017) представлено эмпирическое исследование по прогнозированию банкротства банка с помощью двух подходов: машинного обучения и двух традиционных статистических подходов. [14] Авторы заметили, что методы машинного обучения, Artificial Neural Network (ANN) и k-Nearest Neighbors (k-NN) работают чуть более эффективно, чем традиционные методы. Однако разница в точности прогнозирования между методами ANN и k-NN и традиционным методом логистической регрессии не очень велика. Кроме того, авторы заметили, что SVM (Support Vector Machines) работает не лучше, чем традиционные методы.

# **3. Описание рынка страхования в России за период с 2011 по 2018 года**

С 2011 года количество страховых организаций сократилось почти втрое. Так, в 2012 году рынок потерял 103 участника, а в 2013 году – 23. Такая тенденция продолжается до сих пор: так за 2017 год по данным ЦБ количество страховых организаций сократилось на 33, а годом ранее - на 78. При чем большая часть из них покинула рынок из-за нарушения законодательства, и только меньшая часть - добровольно, вследствие отказа от лицензии. В итоге на конец 2018 года на рынке осталось 199 страховых компаний (рис. 1). Уменьшается в основном доля небольших страховых компаний с уставным капиталом меньше 150 млн рублей, в следствие чего концентрация первых 20 крупных участников рынка по взносам увеличилась до 79,9%.

Рис. 1 - Количество страховых компаний 2011-2018г.

Рис.2 - статистика отзывов лицензий

Так с каждым годом страховым компаниям все сложнее и сложнее удержаться на рынке, это связано так же с ужесточающимся с каждым годом законодательством. Еще десять лет назад минимальный уставной капитал по закону приравнивался к 30 млн, а на 2018 год планка поднялась до 120 млн рублей. При чем совокупный уставной капитал от этого только уменьшился, так на 4-ый квартал 2018 года совокупный уставной капитал равнялся 200,0 млрд рублей, годом ранее – 204,8 млрд рублей, а на конец 2016 года – 216,4 млрд рублей [15, 16].

Многие компании не укладываются в нормативные требования, накладываемые государством, в следствие чего у них отзывают лицензии. А некоторые сами уходят с рынка, добровольно, но это чревато тем, что перед своим уходом они начинают «пылесосить рынок» и собирать страховые премии по низким тарифам, не собираясь ничего потом выплачивать. В следствие чего своим уходом они сильно дестабилизируют экономику и ставят под угрозу благополучие множества физических и юридических лиц.

Несмотря на то, что компании уходят с рынка, и, казалось бы, страховой бизнес не выгоден, страховые премии из года в год растут, и с 2011 года выросли более чем в 2 раза, тогда как процент выплат уменьшился с 45,69% до 35,35%, в расчете на одну компанию страховых взносов становится больше, учитывая отрицательную динамику изменения количества компаний (рис. 3). Но если смотреть динамику страховых поступлений относительно ВВП России, то рост окажется не таким значительным (рис. 4).

Рис. - 3 страховые премии и выплаты

Рис. - 4 Отношение страховых премий к ВВП

Ухудшение макроэкономической ситуации в стране создает финансовые трудности для всех участников хозяйственных отношений, оказывающие негативное влияние на экономическое воспроизводство и сложившуюся среду жизнедеятельности общества в целом. Страховая защита приходит на помощь в такие ситуации, тем самым стабилизируя экономику и развивая рыночные отношения. Так как страховые компании берут на себя риски, возникающие в ходе ведения финансовой деятельности различных организаций, то это стимулирует и деловую активность субъектов. [17] Но в случае финансового кризиса 2007 года пострадали и сами страховщики, которые покрывали риски по ипотечным кредитам, риски финансовых операций или предоставляли гарантированные займы. Значительный убыток в их балансе возник из-за единовременного увеличения уровня выплачиваемых компенсаций. [18] А в 2012 году в результате ошибочной политики государства произошел кризис на рынке ОСАГО, приведший к признанию его нерентабельным и повышению тарифов по ОСАГО на 23-30 % (указание ЦБ РФ от 19.09.2014 № 3384-У). [19][20] Останется ли копания на рынке или нет, зависит от того, насколько сбалансирован совокупный портфель направлений ее деятельности. То есть в случае начала кризиса в одной отросли, это не должно перекрыть прибыль от других направлений ее деятельности. Поэтому необходимо учитывать вид предлагаемых страховыми компаниями страховых услуг и их разнообразие. (Рис. 5)

Рис.5 - Структура страхового рынка по объему премий премии на 2018 год

# **4. Данные для эмпирического исследования**

В дипломной работе за дату дефолта страховой компании взята дата отзыва лицензии центральным банком, являющимся регулятором. В следствие отзыва лицензии прекращается вся страховая деятельность организации под ее регистрационным номером, и начинается процесс ликвидации компании, если ещё не был начат ранее.

Согласно Статье федерального закона (Закон РФ от 27.11.1992 N 4015-1 (ред. от 28.11.2018) "Об организации страхового дела в Российской Федерации") [21] отзыв лицензии страховой компании происходит по решению суда либо органа страхового надзора, в том числе по собственному заявлению страховщика об отказе от продолжения ведения страховой деятельности. По закону основаниями для отзыва лицензии являются:

* инвестиции средств в неразрешенные активы или в случае, когда инвестированные средства превышает 20% от величины страховых резервов или капитала компании;
* соотношение собственных средств и принятых обязательств более чем на 20% отличается от минимально допустимого значения норматива
* размер уставного капитала страховщика ниже минимального значения
* нарушение правил формирования страховых резервов, повлекшее занижение суммы страховых резервов более чем на 20%
* неоднократное непредставление или нарушение сроков представления отчетности или установление фактов существенной недостоверности отчетности
* если субъект в течение года не осуществлял страховую деятельность

Так же регулятор может вынести решение об отзыве лицензии в следствие недостатка ликвидности и платёжеспособности страховой компании, а именно не способности полного и своевременного выполнения обязательств перед клиентами. Видимым результатом данных проблем являются задержки страховых выплат и уклонение от возмещения ущерба под любым предлогом.

Чтобы определить жизнеспособность страховщика, регулирующие органы полагаются на финансовые отчеты, подготовленные в соответствии с установленными законом принципами бухгалтерского учета. Для исследования было решено использовать годовые данные финансовой отчетности в виде интервальных значений на конец года в период с 2012 по 2017 гг., а именно: бухгалтерский баланс страховщика (Форма 1) и отчет о финансовых результатах страховщика (Форма 2). Для этого были сравнены базы данных СПАРК [22] и архивы ЦБ [23] с бухгалтерскими балансами по российским страховым компаний. К достоинствам СПАРК относительно ресурсов ЦБ относится то, что все показатели уже объединены в сводные таблицы со всеми компаниями, но количество самих показателей очень ограничено и представлены несколькими абсолютными, а все остальные относительные, поэтому вычислить новые на основе данных является невозможным. Так же наблюдений с 2012 по 2017 год в базе ЦБ больше, чем в СПАРК, но отчетный период в СПАРК представлен с 1999 года. Огромный недостаток базы ЦБ в том, что вся финансовая отчетность страховых компаний представлена в виде отдельных excel-файлов или zip-архивов с отдельными xml-файлами, поэтому на собирание сводных таблиц с данными уходит не мало времени. Так же минус исходников с годовой финансовой отчетности в том, что формат предоставления отчетности с 2017 года изменился в связи с началом перехода на МСФО, а за 2014 год в архиве выложена только отчетность для формата 1С, что так же создает трудности для сбора данных. Зато база ЦБ является открытым источником данных, в отличии от СПАРК, поэтому любой может повторить вычисления, произведенные в данном исследовании. В данной работе было принято решение остановиться на архивах ЦБ в следствие наибольшей полноты данных.

Для каждого отчетного года данные скачивались из архива ЦБ единым zip-архивом, содержащим файлы с расширением xml. Для объединения всех наблюдений по всем годам в одну сводную таблицу был разработан алгоритм и написана программа на языке программирования Python, которая подвергает парсингу все xml-файлы с данными по первой и второй форме отчетности. Результатом проведенной работы стала итоговая база, содержащая 94 финансовых показателя и 1809 наблюдений по 468 компаниям за период с 2012 по 2016 гг. Так же было получено еще 450 наблюдений за 2011 отчетный период только для тех компаний, для которых не наступил дефолт в течение 2012 года (данные дополнительные наблюдения пригодились для подсчета относительных показателей, отвечающих за динамику изменения, а при построении модели эти наблюдения не участвуют в расчетах). Для 2017 года была сформирована отдельная таблица с 216 наблюдениями и новым форматом колонок вследствие начала перехода на международный формат отчетности Данные в таблицах заполнены по-разному, так, к примеру, показатели «баланса», «активов», «капитала» и «обязательств» совсем не содержат пропусков, в отличие от 13 показателей, заполненных менее чем на 10%. Использование квартальных отчетов представляется нецелесообразным в силу ничтожного количества дефолтов относительно всего количества наблюдений, то есть количество дефолтных значений составляет менее 1% от всей выборки.

Данные по дефолтам страховых компаний были получены из сводной таблицы со сведениями по отзывам лицензий из интернет-источника insur-info.ru, который отслеживает новости на сайте банка России. Но сопоставляя эти данные и реестр ЦБ действующих компаний на данный момент, обнаружились пропуски в данных по ряду компаний. Информацию по этим компаниям пришлось восстанавливать из других интернет-ресурсов, в частности с сайта banki.ru, который опять же ссылается на приказы регулятора. Страховой компании в каждом отчетном периоде проставлялось ее состояние: «1» - у страховой была отозвана лицензия в течение следующего календарного года; «0» - отзыва не было, и страховая еще как минимум год будет осуществлять свою деятельность.

За 6-летний период (2013-2018 гг.) было зафиксировано 267 фактов отзыва лицензии по различным причинам. Но так как для некоторых компаний причиной отзыва являлось непредоставление финансовой отчетности либо недостоверность данных, то по некоторым компаниям в преддефолтные годы отсутствуют данные, особенно ближе к сроку закрытия. Поэтому количество проставленных флагов дефолта в выборке гораздо меньше и достигает 226. Но даже так наша база данных остается достаточно сбалансированной, так как доля наблюдений по компаниям с отозванными лицензиями примерно равна 10%.

Выборка была разделена на обучающую и тестовую. По первой, включающей наблюдения с 2012 по 2016 гг., строились модели, по наблюдениям за 2017 год оценивалась прогнозная сила. (Табл. 1)

Табл. 1

|  |  |  |
| --- | --- | --- |
|  | **Количество наблюдений** | **Доля дефолтов, %** |
| **Вся выборка**  **(2012-2017 гг.)** | 2025 | 11.16% |
| **Обучающая выборка (2012-2016 гг.)** | 1809 | 10.89% |
| **Тестовая выборка (2017 г.)** | 216 | 13.42% |

## **4.1. Описание логит-регрессии**

Для оценивания вероятности дефолта была использована logit-модель бинарного выбора, которая считается одной из наиболее часто применяемых моделей параметрического прогнозирования дефолтов как в научной литературе, так и в банковском и страховом регулировании и надзоре. Логит-модель основана на биномиальной регрессии и строится на оценке вероятности дефолта P (Z), где переменная Z является линейной комбинацией показателей Xi с коэффициентами регрессии bi:

Связь между зависимой переменной, вероятностью дефолта (default), и объясняющими переменными выражается следующим образом:

При этом

В нашем анализе мы рассматриваем факторы (Xi), вычисленные на основе финансовой отчетности, как объясняющие переменные. Оценив коэффициенты Logit-модели, мы получим оценку каждой страховой компании Za:

В результате мы предсказываем вероятность дефолта для каждой страховой компании:

Где - значение объясняемой бинарной переменной дефолта, вектор является -ой строкой матрицы с наблюдениями и объясняющими факторами, а – вектор столбец коэффициентов регрессионной модели. Параметры оценены с помощью функции максимального правдоподобия:

Предполагаемая вероятность дефолта позволяет соотнести каждого страховщика с определенным классом риска. Для этого устанавливается пороговое значение вероятности дефолта P \* для разделения страховых компаний на две группы. Если предполагаемая вероятность дефолта больше, чем P \*, тогда страховщик будет считаться банкротом, и наоборот, если предполагаемая вероятность дефолта ниже, чем P \*, он будет считаться действующим. В большинстве предыдущих работ на эту тему компании считаются по умолчанию преддефолтными, если их вероятность больше или равна 0,5.

## **4.2. Метрика оценивания модели**

Качество моделей классификации проверяется с использованием Receiver Operating Curve (ROC) и коэффициента GINI. Вэбб (2002) определяет ROC как кривую на графике с вертикальной осью true positive rate (доля наблюдений от общего количества, верно классифицированных как дефолтные) и с false positive rate (доля наблюдений от общего количества, не являющиеся дефолтными, ошибочно классифицированных как дефолтные) на горизонтальной оси. [24] Все кривые ROC проходят через точки (0,0) и (1,1), и по мере увеличения расстояния кривая перемещается в верхний левый угол (если направление зависимости выбрано не обратное). Идеальная модель должна обнаруживать 100% дефолтов и иметь 0% ложных срабатываний. ROC в случае идеальной модели характеризуется изогнутой кривой, проходящей через координаты (0,0) - (0,1) - (1,1). (Рис. 6)

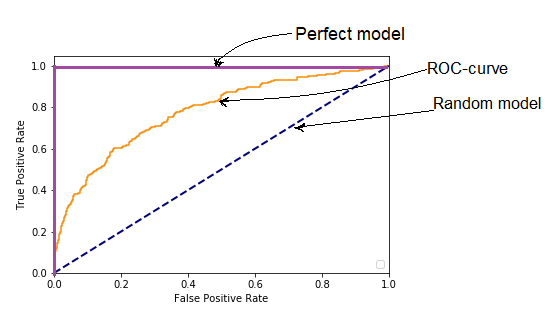


Рис. 6

Разные модели дают разные ROC, характеризующие производительность модели. Производительность определяется как площадь под кривой и обычно обозначается как коэффициент ROC\_AUC. Отсюда следует, что идеальная модель имеет площадь под кривой ROC\_AUC = 1. Для коэффициента GINI, который является площадью между кривой Лоренца и прямой, проходящей через точки (0,0) - (1,1) и являющейся ROC кривой для модели случайного угадывания, справедливо соотношение

## **4.3. Анализ финансовых переменных**

Бухгалтерский баланс страховщика (Форма 1) и отчет о финансовых результатах страховщика (Форма 2) российской бухгалтерской отчетности страховых компаний включает порядка 946 финансовых показателей. Все показатели содержат различное количество наблюдений: так, к примеру, показатели, отвечающие за количество активов, обязательств и капитала, заполнены на 100%, в отличии от 21 финансового показателя, заполненного менее чем на 10%. Изначально планировалось оставить только 30 показателей с наибольшим количеством наблюдений, но в ходе исследования выяснилось, что выкинутые показатели участвуют в вычислении относительных финансовых показателей. А также, если по наблюдению был замечен пропуск в данных, то это не означает отсутствие информации, а лишь показывает, что значение данного показателя равнялось нулю. В связи с этим по всем показателям пропуски в данных были заполнены нулями.

При создании модели предпочтение отдавалось относительным финансовым показателям и небольшому количеству абсолютных, т. к. они легко поддаются экономической интерпретации и позволяют предположить направление и вид их зависимости от объясняемой переменной. Отбор и вычисление объясняющих переменных был осуществлен на основе опубликованных методик. [ 25, 26 ].

Предполагается, что в изначальных сырых данных нет опечаток и ошибок ввода, причем все основные абсолютные переменные, такие как количество активов, капитала и сумма собранных премий за отчетный период были перепроверены, а именно сравнивались по базам ЦБ и СПАРК, и расхождений не было обнаружено. По показателям, отвечающим за расходы, знаки были приведены к единому формату. Но так как регрессионный анализ чувствителен к выбросам и неоднородности данных, была произведена нормализация значений итоговых относительных показателей.

Все финансовые показатели были разделены на группы: показатели финансовой устойчивости, показатели рентабельности, показатели убыточности, показатели оценки перестраховочных операций и инвестиционной деятельности, показатели оценки платежеспособности страховой компании и оценки ее ликвидности и показатели страховой деятельности и доли рынка. Всего 37 различных финансовых показателей разделены на 8 групп.

### **4.3.1. Показатели финансовой устойчивости (financial stability)**

Наиболее глубокую оценку финансового состояния страховой организации дают показатели финансовой устойчивости, но единого мнения по поводу их оценки в различных работах не существует, как и единого определения. Так, А.Д. Шеремет дает определение финансовой устойчивости как «определенное состояние счетов предприятия, гарантирующее его постоянную платежеспособность». [27] Авторы Копбаева Л.Т. и Данилова Н.Л. считают: «Финансовая устойчивость — это стабильность финансового положения предприятия, обеспечиваемая достаточной долей собственного капитала в составе источников финансирования». [28] [29] В самом общем виде финансовую устойчивость И.В. Пискунова, Буркальцева Д.Д., Блажевич О.Г. и Чередниченко М.С. охарактеризовывают как отражение стабильного превышения доходов над расходами предприятия, которое дает возможность эффективно использовать его денежные средства. [30] [31] Куликов С.В. дал свое определение финансовой устойчивости страховой компании, по его мнению – «это способность страховщика выполнять свои обязательства перед другими субъектами хозяйствования при любой неблагоприятной ситуации как в настоящем, так и в будущем». [32]

Табл. 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Обозна-чение** | **Название переменной** | **Формула** | **Формула по бух. Отчетности 2012-2016 гг.** | **Формула по бух. Отчетности с 2017 г.** |
| fs1 | Уровень обеспеченности собственным капиталом | Собственный капитал / (Обязательства + Собственный капитал) | 1\_2100/(1\_2200+1\_2100)[[1]](#footnote-1) | 1\_51 / (1\_40 + 1\_51) |
| fs2 | Уровень покрытия страховых резервов собственным капиталом | Собственный капитал / Страховые резервы по всем видам страхования (без учета перестраховщиков) | 1\_2100/(1\_2210 + 1\_2220 - 1\_1230 - 1\_1240) | 1\_51 / (1\_30 + 1\_33 – 1\_9 – 1\_11 ) |
| fs3 | Уровень долговой нагрузки | Обязательства (без учета страховых резервов и доходов будущих периодов) / (Собственный капитал + обязательства) | (1\_2200 - 1\_2210 - 1\_2220 - 1\_2280) / 1\_2000 | (1\_40 – 1\_30 – 1\_33 ) / 1\_52 |
| fs4 | Коэффициент соотношения заемных и собственных средств | обязательства / собственные средства заемщика | 1\_2200/1\_2100 | 1\_40 / 1\_51 |
| fs5 | Уровень страховых резервов | страховые резервы / сумма активов | 1\_2220/1\_1300 | 1\_33 / 1\_23 |
| fs6 | Соотношение суммы страховых премий и страховых резервов | страховые премии по всем видам страхования /страховые резервы | (2\_2100 + 2\_1100) / (1\_2220 + 1\_2210) | (2\_1 + 2\_8) / (1\_33 +1\_30) |

(fs1) Уровень обеспеченности собственным капиталом: рассчитывается как отношение собственного капитала к сумме обязательств и собственного капитала и представляет собой ключевой показатель финансовой стабильности страховщика. Ожидается положительная связь между этой переменной и финансовым состоянием страховщика и отрицательное влияние на вероятность обанкротиться, поскольку финансовая стабильность страховщика является важным эталоном для потенциальных клиентов.

(fs2) Уровень покрытия страховых резервов собственным капиталом: данный показатель характеризует способность компании покрыть обязательства по страховой деятельности за счет собственного каптала. Ожидается его положительное влияние на уровень финансовой устойчивости и негативное – на вероятность дефолта.

(fs3) Уровень долговой нагрузки: обратно пропорционален предыдущему показателю, следовательно, ожидается его положительное влияние на вероятность дефолта. Данный показатель показывает то, насколько страховщик зависит от внешних источников финансирования.

(fs4) Коэффициент соотношения заемных и собственных средств: страховое плечо определяется соотношением привлеченных средств (технических резервов) к капиталу. Риск страховщика может увеличиться, когда он увеличивает свой финансовый рычаг. Литература по структуре капитала подтверждает, что стоимость фирмы увеличится до оптимальной точки с увеличением кредитного плеча, а затем снижается, если рычаг будет дополнительно увеличен за пределы оптимального уровня. Таким образом, использование за пределами этого оптимального уровня может привести к высокому риску несостоятельности и низкой стоимости фирмы. Кредитное плечо используется как индикатор агрессивности продаж фирмы и считается, что более высокий операционный рычаг приводит к снижению платежеспособности страховщиков. Финансовый рычаг может измеряться отношением общего долга к собственному капиталу, или как отношение суммы чистых премий и собственного капитала. Это соотношение показывает степень, в которой бизнес использует заемные деньги, и отражает потенциальное влияние дефицита технических резервов на капитал в случае непредвиденных потерь.

(fs5) Уровень страховых резервов: уровень страховых резервов вычисляется как доля страховых резервов в суммарных активах компании и чем больше значение данного показателя, тем больше шансов, что страховщик обанкротится.

(fs6) Соотношение суммы страховых премий и страховых резервов: страховые резервы формируются за счет страховых премий и используются в случае наступления страхового случая для выплат по договору страхования. Следовательно, слишком низкое значение соотношения показывает, что страховщик набрал много плохих договоров и ожидает большие выплаты по ним, а слишком малое значение данного показателя говорит о том, что страховщик собирается наоборот уклоняться от выплат.

### **4.3.2. Показатели Рентабельности (profitability)**

Показатели рентабельности характеризуют результативность и эффективность деятельности страховой компании. При помощи показателей рентабельности оцениваются прибыль и убытки деятельности страховой компании. Рентабельность является одной из важнейших задач финансового менеджмента, потому что одной из его целей является максимизация благосостояния владельца, и, в данном случае, прибыльность является очень важным определяющим фактором эффективности.

Доходность компании, занимающейся страхованием, критически зависит от ее операционной и финансовой деятельности. Операционная деятельность состоит из страховых операций: заключение новых договоров и обслуживание существующих. Финансовая деятельность заключается в инвестировании страховых премий. Доходы от операционной деятельности проистекают из разницы между доходом от премий и общей стоимостью страхования и операций, тогда как доходы от финансовой деятельности проистекают из разницы между фактическим доходом от инвестиций и доходом, полученным со сборов.

Показатель нормы доходности чистых активов измеряет рентабельность чистых активов страховой компании, а также служит для оценки стоимости долгового капитала. [33] Часто рентабельность вычисляется как отношение прибыли до вычета процентов и налогов к общим активам, но в данном исследовании мы возьмем в виде переменных, отвечающих за рентабельность компании, целый список из 6 факторов. (Табл. 3)

Табл. 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Обозна-чение** | **Название переменной** | **Формула** | **Формула по бух. Отчетности 2012-2016 гг.** | **Формула по бух. Отчетности с 2017 г.** |
| pr1 | Рентабельность страховой и финансово-хозяйственной деятельности | Прибыль или убыток до налогообложения / Сумма доходов страховой компании | (2\_3400 ) / (2\_1100 + 2\_1200 + 2\_1300 + 2\_1700 + 2\_1800 +2\_2110 + 2\_2700 + 2\_2800 +2\_2910 + 2\_2920 + 2\_3200 +2\_3300) | (2\_30) / (2\_1.1 + 2\_5 + 2.6 + 2\_8.1 + 2\_22 + 2\_12 + 2\_13 + 2\_27 + 2\_28) |
| pr2 | Рентабельность собственного капитала | Прибыль или убыток до налогообложения / Собственный капитал | 2\_3400 / 1\_2100 | 2\_30 / 1\_51 |
| pr3 | Чистая рентабельность собственного капитала | Чистая прибыль / Собственный капитал | 2\_3000 / 1\_2100 | 2\_54 / 1\_51 |
| pr4 | Чистая рентабельность страховой деятельности | Чистая прибыль (убыток) / Сумма страховых премий | 2\_3000 / (2\_1110 + 2\_2110) | 2\_54 / ( 2\_1.1 + 2\_8.1) |
| pr5 | ROA, чистая рентабельность активов | Чистая прибыль (убыток) / Активы | 2\_3000 / 1\_1300 | 2\_54 / 2\_8.1 |
| pr6 | ROA, рентабельность активов | Прибыль или убыток до налогообложения / активы | 2\_3400 / 1\_1300 | 2\_30 / 2\_8.1 |

(pr1 и pr4) Рентабельность страховой и финансово-хозяйственной деятельности и чистая рентабельность страховой деятельности: данные показатели воздействуют на нашу объясняемую переменную u-образно.

(pr2/pr3) ROE, [чистая] рентабельность собственного капитала: в корпоративных финансах рентабельность собственного капитала (ROE) является мерой прибыльности бизнеса по отношению к собственному капиталу, также известному как чистые активы или активы за вычетом обязательств. ROE - это показатель того, насколько хорошо компания использует инвестиции для увеличения прибыли. Коэффициент рентабельности собственного капитала вычисляется, как отношение прибыли страховой компании к собственному капиталу, показывает полученную прибыль на единицу капитала. Следовательно, большое значение данного показателя говорит о том, что деятельность страховой компании является рентабельной, а значит и уменьшается вероятность отзыва лицензии.

(pr5/pr6) ROA, [чистая] рентабельность активов: рентабельность активов (ROA) показывает процент прибыльности активов компании, то есть отражает в какой степени страховая компания эффективно использует свои активы. Это число говорит о том, что компания может сделать с тем, что она имеет, то есть сколько долларов прибыли страховщик получают от каждого доллара активов, которого он контролирует. Данное соотношение является важным фактором в финансовом анализе при сравнении конкурирующих компаний в одной отрасли. Чем больше значение данных показателей, тем меньше вероятность дефолта.

### **4.3.3 Показатели убыточности (unprofitability)**

Показатели убыточности, использованные в данном исследовании представлены в таблице 4.

Табл. 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Обозна-чение** | **Название переменной** | **Формула** | **Формула по бух. Отчетности 2012-2016 гг.** | **Формула по бух. Отчетности с 2017 г.** |
| unpr1 | Показатель убыточности | Расходы по страховой деятельности / Страховые премии | (2\_2200 + 2\_1400) /(2\_2100 + 2\_1100) | (2\_9 + 2\_2) / (2\_8 + 2\_1) |
| unpr2 | Показатель уровня расходов (расходы на ведение дела) | Расходы страховой компании (прочие) / Страховые премии | – (2\_1600 +2\_1800 + 2\_1700 + 2\_2220 + 2\_2600 + 2\_2920 + 2\_2910 + 2\_3100 + 2\_3300 + 2\_3200) / (2\_1100 + 2\_2100 ) | – (2\_4 + 2\_5 +2\_6 + 2\_9.2 + 2\_10 + 2\_12 + 2\_13 + 2\_23 +2\_27 +2\_28) / (2\_8 + 2\_1) |
| unpr3 | Комбинированный коэффициент убыточности | Расходы страховой компании (все) / Страховые премии | – (2\_1400 + 2\_1500 + 2\_1600 + 2\_2200 + 2\_2600 + 2\_3100) / (2\_2100 + 2\_1100) | – (2\_2 + 2\_3 + 2\_4 + 2\_9 + 2\_10 + 2\_23) / (2\_8 + 2\_1) |

(unpr1) Показатель убыточности: показатель убыточности страховщика отражает уровень выплат относительно премий компании без учета премий, переданных в перестрахование. Данный показатель определяет насколько объем заработанных страховых премий покрывает расходы компании по страховым выплатам страховой и уровень убыточности собственных страховых операций. Существует прямая зависимость между показателем убыточности и вероятностью дефолта.

(unpr2) Показатель уровня расходов (расходы на ведение дела): данный показатель характеризует неэффективность ведения операционной деятельности и показывает отрицательное влияние на финансовое состояние компании.

(unpr3) Комбинированный коэффициент убыточности: комбинированный коэффициент убыточности отражает общую эффективность деятельности страховщика и прямо пропорционально влияет на вероятность обанкротиться.

### **4.3.4. Показатели оценки перестраховочных операций и инвестиционной деятельности (reinsurance operations and investment activity)**

Показатели оценки перестраховочных операций и инвестиционной деятельности, использованные в данном исследовании, представлены в таблице 5.

Табл. 5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Обозна-чение** | **Название переменной** | **Формула** | **Формула по бух. Отчетности 2012-2016 гг.** | **Формула по бух. Отчетности с 2017 г.** |
| re1 | Доля перестраховщиков в  страховых резервах  (кроме жизни) | Доля перестраховщиков в страховых резервах / Страховые резервы | 1\_1240 / 1\_2220 | 1\_11 / 1\_33 |
| re2 | Доля страховых премий, переданных в перестрахование | Страховые премии, переданные перестраховщикам / Страховые премии - всего | (-(2\_2120 + 2\_1120) / (2\_1110 + 2\_2110)) | (2\_8.2 + 2\_1.2) / (2\_1.1 + 2\_8.1) |
| re3 | Доля перестраховщиков в выплатах | Страховые выплаты, переданные перестраховщикам / Страховые выплаты - всего | (2\_1420 + 2\_2230)/(2\_1400 + 2\_2210) | (2\_8.2 + 2\_1.2) / (1\_33 – 1\_11 |
| inv | Уровень покрытия инвестиционными активами страховых резервов-нетто (кроме страхования жизни) | (Финансовые вложения + Денежные средства) / Страховые резервы без перестраховщиков | (1\_1140 + 1\_1270 ) / ( 1\_2220 – 1\_1240) | (1\_1 + 1\_2 + 1\_3 + 1\_4 + 1\_5) / (1\_33 – 1\_11) |

(re1, re2) Перестрахование: зависимость перестрахования рассчитывается как отношение брутто-премий, удержанных в перестраховании, к совокупным сборам или как доля перестраховщиков в страховых резервах. Страховые компании перестраховывают определенную сумму риска, гарантированную для снижения риска банкротства в случае высоких потерь. Ожидается, что если компания слишком сильно полагается на перестраховщиков либо совсем ими пренебрегает, то это негативно сказывается на ее деятельности.

(re3) Доля перестраховщиков в выплатах: показатель, характеризующий насколько хорошо страховая компания выявила плохие договора и перестраховала их. Следовательно, чем больше выплат было сделано со стороны перестраховщиков, тем меньше вероятность дефолта страховой.

(inv) Уровень покрытия инвестиционными активами страховых резервов-нетто: уровень покрытия инвестициями показывает сбалансированность инвестиционных решений. Таким образом, данный показатель становится критически важным для финансовой надежности страховщика. Многие исследователи считают, что уровень покрытия инвестиционными активами страховых резервов отрицательно коррелирует с показателем несостоятельности. Коэффициент инвестиций рассчитывается путем деления инвестиций на общие резервы. Поскольку инвестиции приносят инвестиционный доход, то ожидается положительное влияние этой переменной на финансовые показатели. Но слишком большие финансовые вложения и маленькие резервы могут тоже привести к дефолту фирмы. Следовательно, существует нелинейная связь между объясняемой и объясняющей переменными.

### **4.3.5 Показатели оценки платежеспособности страховой компании и оценки ее ликвидности (liquidity and solvency)**

Платежеспособность в области финансов или бизнеса - это показатель, при котором текущие активы физического или юридического лица превышают текущие обязательства этого физического или юридического лица. Платежеспособность также можно охарактеризовать как способность корпорации вовремя погашать свои краткосрочные и долгосрочные обязательства. [34] Тогда как в обратную сторону, неплатежеспособность может привести к банкротству компании. По Федеральному закону от 28.04.2009 N 73-ФЗ неплатежеспособность – это «прекращение исполнения должником части денежных обязательств или обязанностей по уплате обязательных платежей, вызванное недостаточностью денежных средств. При этом недостаточность денежных средств предполагается, если не доказано иное». Платежеспособность является первоочередным показателем для регулирующих органов страховых компаний, основным критерием для держателей страховых полисов, а возврат инвестиций является основной задачей для инвесторов.

Важной подкатегорией страхового регулирования является регулирование платежеспособности. Цель регулирования платежеспособности - защитить страхователей, гарантируя, что страховщик сможет выполнить свои финансовые обязательства в будущем. Чтобы иметь возможность как можно раньше вмешаться и свести к минимуму потенциальные издержки, связанные с финансовым крахом страховщика, регуляторы используют системы раннего предупреждения для выявления компаний с финансовыми проблемами. Прогнозирование платежеспособности должно оставаться постоянным даже в периоды кризисов, а модели прогнозирования должны обеспечивать надежные оценки будущего финансового положения страховщиков. [35]

Другим определяющим финансовым показателем является уровень ликвидности. Ликвидность означает степень, в которой долговые обязательства, поступающие в течение следующих двенадцати месяцев, могут быть выплачены из денежных средств или активов, которые будут переведены в денежные средства. Страховая ликвидность - это способность страховщика выполнять свои непосредственные обязательства перед держателями страховых полисов без увеличения прибыли от андеррайтинга и инвестиционной деятельности и / или ликвидации финансовых активов. Остатки денежных средств и банковских счетов должны быть достаточными для удовлетворения немедленных (неотложных) обязательств по искам, подлежащих оплате, но еще не урегулированных. (Табл. 6)

Табл. 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Обозна-чение** | **Название переменной** | **Формула** | **Формула по бух. Отчетности 2012-2016 гг.** | **Формула по бух. Отчетности с 2017 г.** |
| liq1 | Текущая ликвидность | Ликвидные активы / Обязательства страховой компании | (1\_1140 + 1\_1270) / (1\_2200 – 1\_1230 - 1\_1240 – 1\_2280 ) | (1\_1 + 1\_2 + 1\_3 + 1\_4 + 1\_5 ) / (1\_40 – 1\_9 - 1\_11 ) |
| liq2 | Доля наиболее ликвидных активов в общем объеме активов | Сумма наиболее ликвидных активов / Активы страховой компании | (1\_1140 + 1\_1270) / 1\_1300 | (1\_1 + 1\_2 + 1\_3 + 1\_4 + 1\_5 ) / 1\_23 |
| solv | Текущая платежеспособность | Страховые премии (взносы) / Расходы страховой компании | (2\_1100 + 2\_2100) / (2\_1400 +2\_1600 + 2\_1800 + 2\_2200 +2\_2600 + 2\_2920 + 2\_3100) | (2\_8 + 2\_1) / (2\_2 + 2\_4 + 2\_6 + 2\_9 + 2\_10 + 2\_13 + 2\_23) |

(liq1, liq2) Коэффициенты ликвидности: ликвидность - это способность страховщика погашать обязательства, которые включают операционные расходы и оплату убытков / пособий по страховым полисам, когда это необходимо. Для страховщика денежные потоки (главным образом, премии и инвестиционный доход) и ликвидация активов являются двумя источниками ликвидности. При этом текущий коэффициент ликвидности является важным показателем платежеспособности и обозначает, какую часть обязательств компания сможет погасить единовременно. Коэффициенты ликвидности оказывают положительное влияние на стабильность компании и ее платёжеспособность и отрицательное - на несостоятельность. Но одновременно с этим слишком большая доля наличных денег на счету организации говорит о скором выводе денег из нее либо о нерациональном использовании активов, которых можно было пустить в инвестиции. Следовательно, существует нелинейная зависимость между переменными ликвидности и вероятностью дефолта.

(solv) Текущая платежеспособность: данный показатель рассчитывается как отношение страховых взносов к общим расходам страховщика и показывает способность компании своевременно и полностью выполнять свои обязательства. Чем меньше данный показатель, тем больше вероятность отзыва лицензии предприятия.

### **4.3.6. Показатели страховой деятельности и доли рынка**

Данные показатели были взяты из сводных архивов ЦБ [36], и большинство из них не отображаются в годовой финансовой отчетности, а сдаются по указу ЦБ в отчетности по форме 0420162 "Сведения о деятельности страховщика" с конца октября 2017 года, или по ранее выпущенному приказу Росстата - в форме N 1-СК. (Табл. 7)

Табл. 7

|  |  |
| --- | --- |
| **Обозначение** | **Фактор** |
| prem | Размер премии |
| pay | Размер выплат |
| delta\_prem | Изменение совокупного объема премий |
| delta\_pay | Изменение совокупного объема выплат |
| act | Размер страховой компании (сумма активов) |
| insure\_sum\_new | Страховые суммы по договорам страхования, заключенным в отчетном периоде |
| num\_zayav | Кол-во заявленных страховых случаев |
| num\_ureg | Количество урегулированных страховых случаев |
| num\_ref | Количество отказов страховой выплаты |
| num\_all\_new | Количество заключенных договоров страхования |
| num\_all | Количество действовавших договоров страхования на конец периода |
| num\_ref\_div\_zayav | Доля отказов выплаты относительно заявленных |
| num\_zayav\_div\_all | Доля наступивших страховых случаев относительно количества всех заключенных |
| prem\_div\_insure\_sum\_new | Отношение премий к страховым суммам |
| ca | Капитал |

(prem, act) Размер фирмы (сумма премий или активов): на финансовое состояние любой организации влияют, помимо прочего, размер или общие активы фирмы. Поскольку регуляторы с меньшей вероятностью ликвидируют крупных страховщиков, ожидается, что малые страховщики находятся в более уязвимом положении и больше подвержены вероятности наступления несостоятельности. Размер фирмы влияет на ее финансовые показатели разными способами. Крупные фирмы могут использовать экономию за счет масштаба и тем самым более эффективно вести свою деятельность по сравнению с небольшими фирмами. Так же размер компании может быть определен и чистой премией, которая вычисляется как сумма сборов, полученных страховой компанией после вычета выплат по перестрахованию. Так же переменные, используемые для измерения размера фирмы, включают общую премию, общие допустимые активы, капитал и излишек.

(delta prem) Рост премий: объем премий определяет уровень охвата рынка, а рост премий определяет скорость проникновения на рынке. Хотя рост является важным фактором роста стоимости для всех фирм, он имеет особое значение для страховых компаний. Эффективная работа этих фирм требует значительной экономии от масштаба, создаваемой объемом бизнеса. Без роста страховщик не может получить объем бизнеса, необходимый для обеспечения коллективного объединения страховых рисков в соответствии с законом больших чисел, на котором основываются страховые операции. На внутреннем рынке рост достигается в основном за счет расширения систем дистрибуции и совершенствования технологий. Эмпирические результаты показывают, что если сборы страховой компании увеличиваются, то это значит, что дела у компании идут хорошо, и она развивается, а, следовательно, повышается ее стабильность на рынке, и снижается вероятность стать банкротом.

Но с другой стороны стремительный быстрый рост объема премии является одним из причинно-следственных факторов несостоятельности страховщиков. Если компании слишком одержима ростом, то это может привести к ее банкротству, поскольку она пренебрегает другими важными факторами. Высокий рост премий связан с повышением риска компании, поскольку стратегия агрессивного роста может увеличить риск неплатежеспособности. В частности, чрезмерный рост является проблемой во время экономического спада. Если фирма принимает на себя существенные риски, увеличивая премиальный доход - например, за счет снижения цены по тарифам и привлекая больше клиентов, то она не сможет перенести финансовые последствия в случае неблагоприятного развития в краткосрочной перспективе, хотя такая стратегия может окупиться в долгосрочной перспективе. Это может привести к более высокому риску неплатежеспособности. Таким образом, платежеспособность страховщиков с агрессивными стратегиями роста вероятно ухудшается во время финансового кризиса. Рост премий, включенный в анализ, - это процентное изменение в годовом исчислении чистых премий.

(pay) Размер выплат по договорам страхования: рассчитывается в абсолютном выражении по модулю, как сумма выплат по договорам страхования жизни и иным без договоров перестрахования. Считается, что чем больше выплаты, а, следовательно, и расходы компании, тем больше вероятность дефолта страховщика.

(delta\_pay) Изменение совокупного объема выплат: данный показатель рассчитывается как отношение разницы выплат текущего и прошлого года к текущим выплатам. Предполагается, что стремительный рост выплат по договорам страхования должен отрицательно сказываться на балансе компании и на ее финансовом состоянии, и, следовательно, существует прямая зависимость между ростом выплат и вероятностью дефолта компании.

(insure\_sum\_new) Страховые суммы по договорам страхования, заключенным в отчетном периоде: сумма, которую страховщик обязуется выплатить в случае наступления страхового случая, и она не должна превышать стоимость страхуемого имущества на момент заключения договора. Следовательно, чем больше значение данного показателя, тем больше обязательств взял на себя страховщик и с тем большей вероятностью он обанкротится.

(num\_zayav и num\_ureg) Количество заявленных и урегулированных случаев: данные показатели схожи с суммой выплат, только представлены в количественном виде. Следовательно, предполагаем, что они отрицательно влияют на стабильность компании, и положительно на вероятность дефолта.

(num\_ref) Количество отказов в выплате: когда некоторые страховщики собираются уходить с рынка, то перед своим уходом они зачастую начинают «пылесосить рынок» и собирать страховые премии по низким тарифам, не собираясь ничего потом выплачивать. Поэтому регулятор может вынести решение об отзыве лицензии в следствие задержек страховых выплат и уклонения страховой компании от возмещения ущерба под любым предлогом.

(num\_all\_new и num\_all) Количество заключенных договоров страхования в отчетном периоде и количество действовавших договоров страхования на конец периода: данные показатели характеризуют размер страховой компании, как и упомянутые ранее сумма премий и величина активов. Следовательно, существует обратная зависимость между количеством заключенных договоров и вероятностью дефолта страховщика.

(num\_ref\_div\_zayav) Доля отказов в выплате ко всем заявленным: чем больше данное отношение, там больше вероятность, что регулятор заинтересуется данной страховой компанией, и больше вероятность отзыва лицензии.

(num\_zayav\_div\_all) Доля наступивших страховых случаев относительно количества всех действующих заключенных договоров: большую прибыль получат те компании, в которых аналитики смогут точнее спрогнозировать вероятность наступления страхового случая, и застраховать только тех лиц, для которых прогнозная вероятность выплаты близка к нулю. Следовательно, данный показатель прямо пропорционально влияет на объясняемую переменную.

(prem\_div\_insure\_sum\_new) Отношение премий к страховым суммам: чем больше премия, заработанная страховой компаний, и меньше сумма, которую придется выплачивать в случае наступления страхового случая, тем лучше для компании, а именно, тем больше её прибыль. Из этого следует, что вероятность дефолта убывает с возрастанием данного отношения.

(ca) Капитал: для обеспечения будущего роста страховая компания должна генерировать и поддерживать достаточный капитал для удовлетворения требований регулирующих органов, а также для финансирования своего расширения. Хотя капитал может генерироваться путем выпуска акций в краткосрочной перспективе, компании по страхованию жизни, как и любые другие компании, должны иметь возможность генерировать капитал с помощью прибыльных операций.

### **4.3.7. Обработка данных**

Для построения логит-регрессий было необходимо избавиться от пропусков в данных и бесконечно больших и малых значений. Для этого было рассмотрено два случая: удаление наблюдений с пропусками или заполнение пропусков медианными значениями или нулями, при чем бесконечные значения были заменены на числовые значения, равные минимальному и максимальному значениям переменной по всем наблюдениям. В первом случае мы лишились 321 наблюдения из 2025, по которым отсутствовал бы прогноз, но Gini логит-регрессии и качество модели от этого повышалось бы. Во втором случае при заполнении пропусков средними значение коэффициента Gini падает как по отдельным переменным, так и по всей модели в целом. А в третьем случае, при осмысленном заполнении пропусков, а именно понимая, как были получены значения «Nan», «inf» и «-inf», мы опять же не теряли наблюдения и значение Gini и качество модели уменьшалось относительно первого варианта слабее. Поэтому было принято решение отдать предпочтение третьему способу заполнения пропусков.

Стоит отметить, что регрессионные методы очень чувствительны к выбросам, поэтому было решено сначала нормализовать данные при помощи s-shape кривой, а не удалять выбросы, там самым не уменьшая выборку еще больше, при чем новые значения всех переменных будут лежать на отрезке от 0 до 1 (Рис. 7). Коэффициент Gini по отдельным переменным от этого не изменился, но, исследования показали, что значение Gini модели от всех переменных выросло почти на 10%, что является положительным результатом исследований.

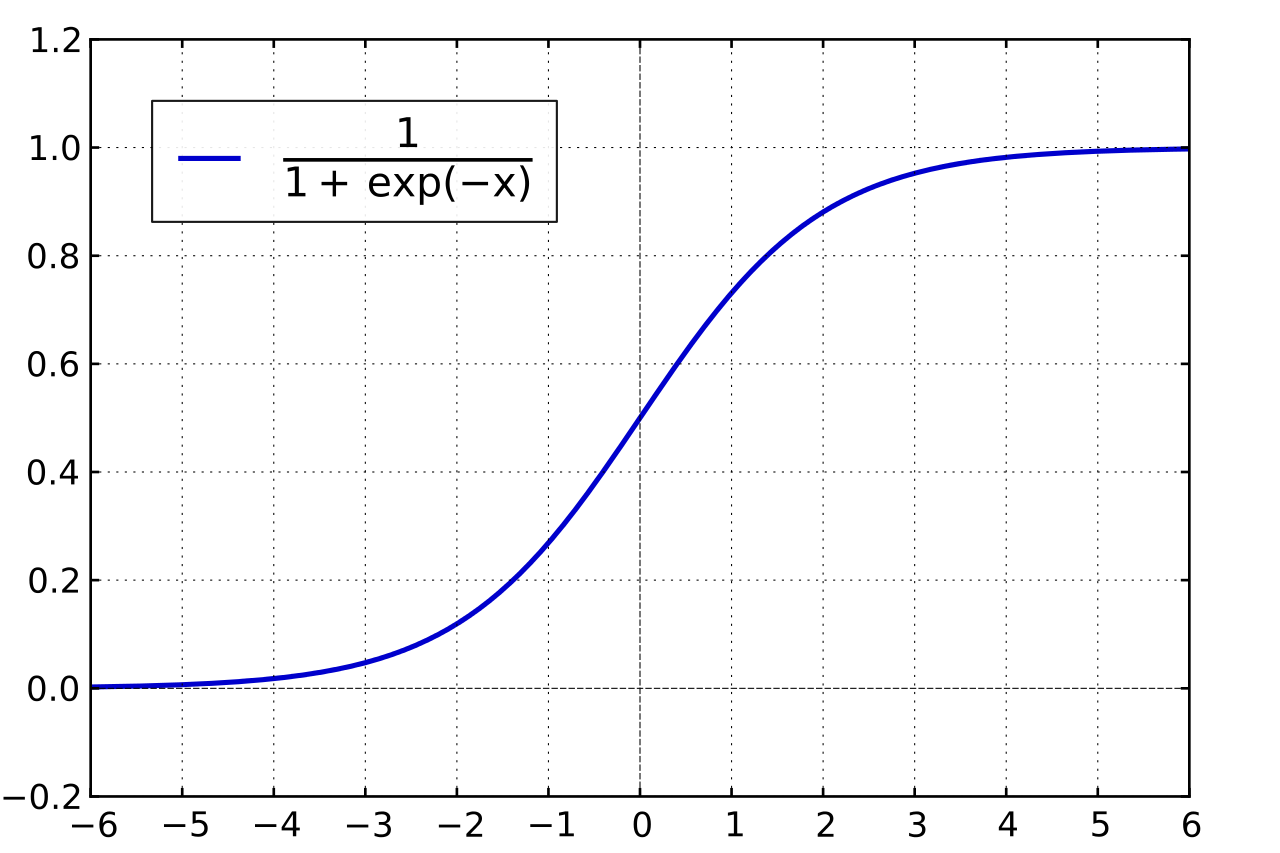


Рис. 7

### **4.3.8. Отбор факторов**

Для всех финансовых переменных была построена roc-кривая, вычислен коэффициент Gini, а также получены значения P-value ANOVA-теста на разделительную способность. Наибольшим Gini (32,6 – 33,4 %) обладают показатели рентабельности и показатели страховой деятельности по количеству заключенных новых договоров и всех действующих на конец отчетного периода. А в тесте на разделительную способность большинство факторов показали значимость на уровне 10%.

Табл. 8

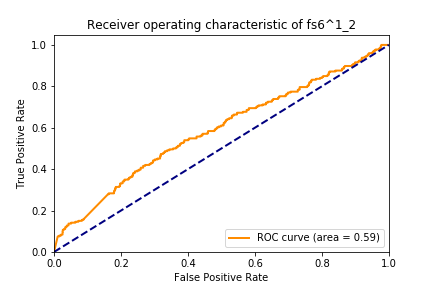
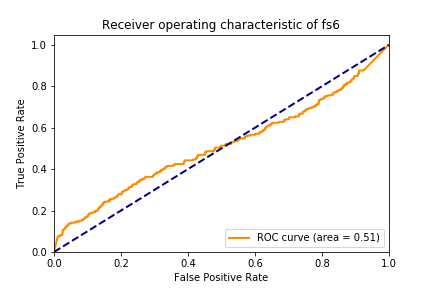
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Переменные** | **Влияние** | **Знак Джини** | **Джини** | **P-value ANOVA-теста** |
| fs1 | -1 | 1 | 0,12 | 0,00 |
| fs2 | -1 | 1 | 0,19 | 0,00 |
| fs3 | 1 | 1 | 0,03 | 0,32 |
| fs4 | 1 | -1 | 0,14 | 0,02 |
| fs5 | 1 | -1 | 0,11 | 0,03 |
| fs6 | u | 1 | 0,02 | 0,02 |
| pr1 | u | -1 | 0,17 | 0,08 |
| pr2 | -1 | -1 | 0,33 | 0,00 |
| pr3 | -1 | -1 | 0,31 | 0,00 |
| pr4 | u | -1 | 0,24 | 0,03 |
| pr5 | -1 | -1 | 0,31 | 0,00 |
| pr6 | -1 | -1 | 0,32 | 0,00 |
| unpr1 | 1 | 1 | 0,08 | 0,05 |
| unpr2 | 1 | 1 | 0,18 | 0,00 |
| unpr3 | 1 | 1 | 0,11 | 0,02 |
| re1 | u | -1 | 0,13 | 0,02 |
| re2 | u | -1 | 0,02 | 0,11 |
| re3 | -1 | -1 | 0,06 | 0,95 |
| inv | u | 1 | 0,06 | 0,00 |
| liq1 | u | -1 | 0,02 | 0,12 |
| liq2 | u | -1 | 0,16 | 0,00 |
| solv | -1 | -1 | 0,14 | 0,00 |
| prem | -1 | -1 | 0,22 | 0,00 |
| pay | 1 | -1 | 0,16 | 0,00 |
| delta\_prem | u | -1 | 0,16 | 0,00 |
| delta\_pay | 1 | -1 | 0,06 | 0,13 |
| act | -1 | -1 | 0,20 | 0,00 |
| insure\_sum\_new | 1 | -1 | 0,23 | 0,08 |
| num\_zayav | 1 | -1 | 0,25 | 0,00 |
| num\_ureg | 1 | -1 | 0,24 | 0,00 |
| num\_ref | 1 | -1 | 0,25 | 0,00 |
| num\_all\_new | -1 | -1 | 0,28 | 0,00 |
| num\_all | -1 | -1 | 0,33 | 0,00 |
| num\_ref\_div\_zayav | 1 | -1 | 0,19 | 0,01 |
| num\_zayav\_div\_all | 1 | -1 | 0,15 | 0,25 |
| prem\_div\_insure\_sum\_new | -1 | -1 | 0,07 | 0,97 |
| ca | -1 | -1 | 0,08 | 0,04 |

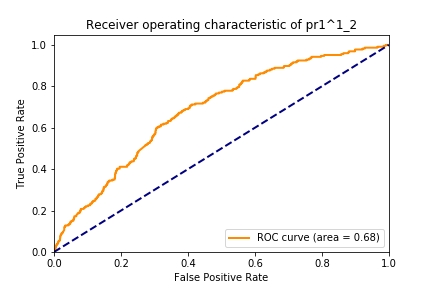
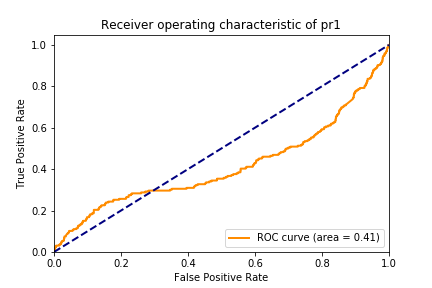
Среди всех переменных были оставлены те, у которых знак коэффициента Джини или форма roc-кривой соответствует нашим экспертным ожиданиям, а именно направлению воздействия на вероятность дефолта страховой. Это может быть нелинейное воздействие (u-образная зависимость), либо прямо пропорциональное, либо обратно пропорциональное. В результате такого сравнения (отбора) из 37 переменных осталось только 25 финансовых переменных, выделенных цветом в таблице. (Табл. 8)

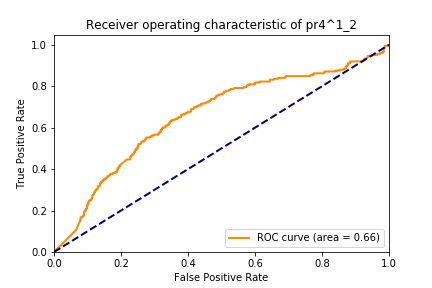
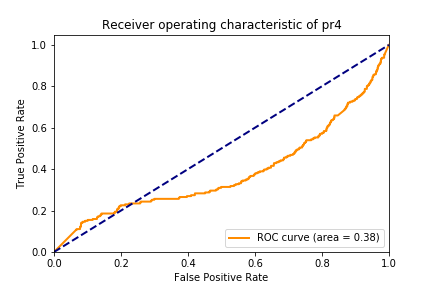
Преобразование нелинейно зависимых переменных

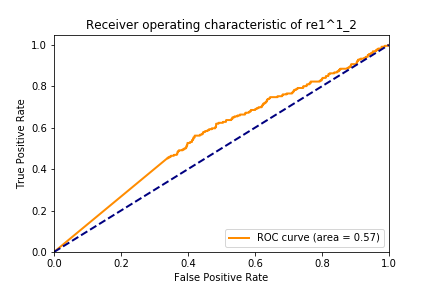
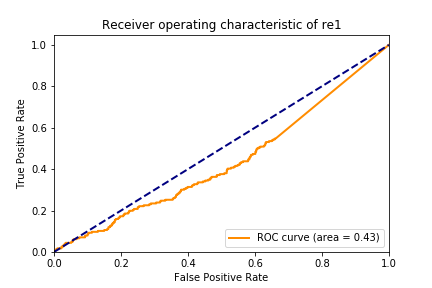
Переменные, у которых нелинейная зависимость с вероятностью дефолта, должны входить в логит-модель нелинейно, чтобы их влияние на объясняющую переменную не меняло своего направления и было постоянным. То есть графически на гистограмме распределения их зависимость лучше всего представляется параболой, уравнение которой представлено линейной комбинацией квадрата данной финансовой переменной и ее изначальным значением.

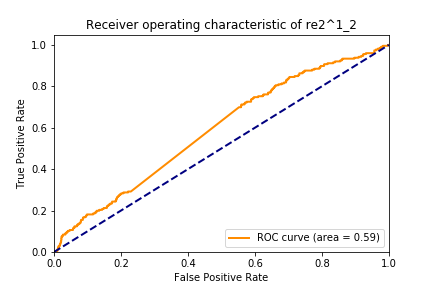
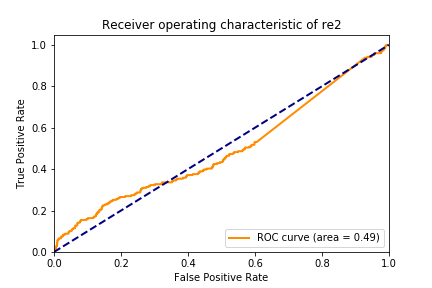
На графике с осями FPR TPR можно заметить, как s-образная roc-кривая пересекает диагональ, а именно кривую безразличия, это говорит о том, что действительно очень много дефолтных наблюдений сконцентрировано в первом и последнем квантиле значений данных финансовых переменных (левая колонка графиков 8-16).

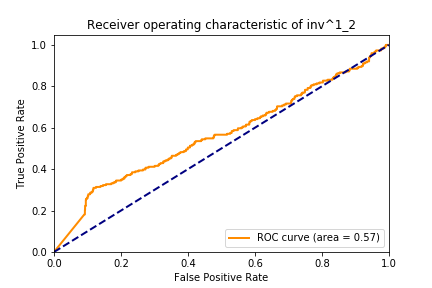
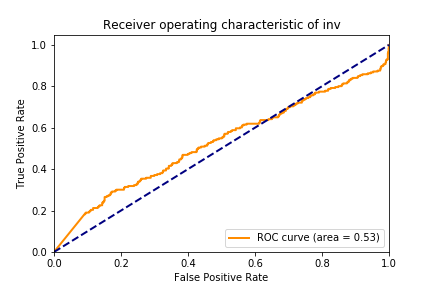


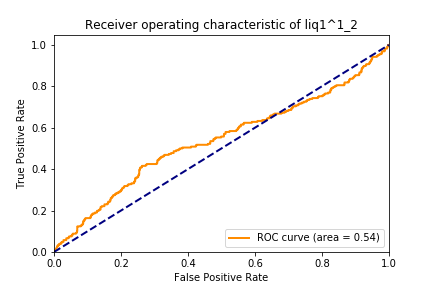
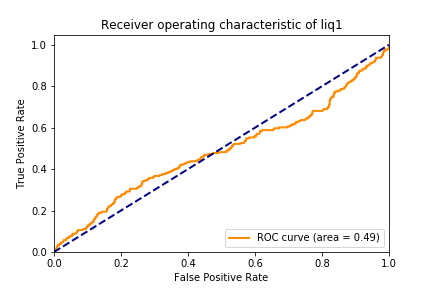


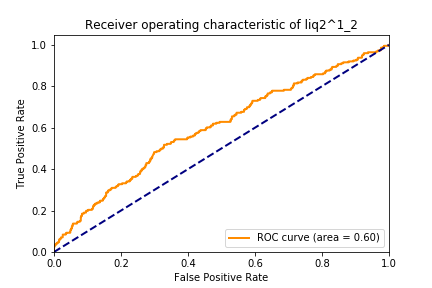
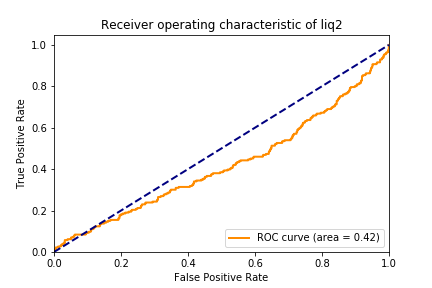












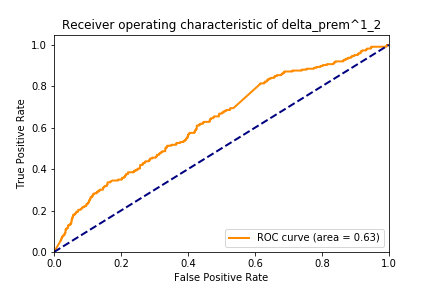
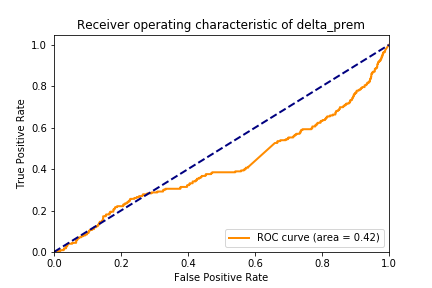


Рис. 8 - 16

Для побора наиболее подходящих коэффициентов и вычисления новой переменной, лежащей в пределах от 0 до 1, была применена логит-регрессия, которой в виде объясняющих переменных на вход подавались попарно сами переменные и их квадраты, а в виде объясняемой переменной были взяты исторические дефолты страховых компаний. Новые предсказанные факторы были добавлены к данным для модели взамен предыдущих факторов. При чем обучение данных моделей происходило на обучающей выборке, а вектора прогнозов строились уже по всей, чтобы избежать неявного попадания вектора дефолтов тестовой выборки в объясняющие переменные.

Судя по ROC-кривым полученных переменных, которые показаны на графиках слева (Рис. 8-16), а также сравнению коэффициентов Джини в абсолютном значении (Табл. 9), можно сделать вывод, что наши гипотезы на счет вида воздействия оправдались и данные преобразования факторов улучшили их предсказательную силу.

Табл. 9

|  |  |  |
| --- | --- | --- |
| **Параметры** | **Старый Gini** | **Новый Gini** |
| **fs6** | 2,2% | 17,4% |
| **pr1** | 17,0% | 36,9% |
| **pr4** | 23,5% | 31,7% |
| **re1** | 13,2% | 13,3% |
| **re2** | 1,5% | 17,7% |
| **inv** | 5,5% | 13,2% |
| **liq1** | 2,3% | 7,7% |
| **liq2** | 15,7% | 20,1% |
| **delta\_prem** | 16,5% | 26,4% |

Проведем исследование корреляции переменных (Табл. 10-12).

Табл. 10

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **fs3** | **pr2** | **pr3** | **pr5** | **pr6** | **unpr1** | **unpr2** | **unpr3** |
| **fs3** | 1,00 | 0,14 | 0,12 | 0,04 | 0,03 | 0,07 | 0,22 | 0,40 |
| **pr2** | 0,14 | 1,00 | 0,98 | 0,84 | 0,85 | 0,15 | 0,33 | 0,08 |
| **pr3** | 0,12 | 0,98 | 1,00 | 0,86 | 0,84 | 0,14 | 0,33 | 0,08 |
| **pr5** | 0,04 | 0,84 | 0,86 | 1,00 | 0,98 | 0,13 | 0,34 | 0,02 |
| **pr6** | 0,03 | 0,85 | 0,84 | 0,98 | 1,00 | 0,14 | 0,34 | 0,02 |
| **unpr1** | 0,07 | 0,15 | 0,14 | 0,13 | 0,14 | 1,00 | 0,04 | 0,29 |
| **unpr2** | 0,22 | 0,33 | 0,33 | 0,34 | 0,34 | 0,04 | 1,00 | 0,30 |
| **unpr3** | 0,40 | 0,08 | 0,08 | 0,02 | 0,02 | 0,29 | 0,30 | 1,00 |
| **re3** | 0,03 | 0,04 | 0,04 | 0,00 | 0,00 | 0,09 | 0,09 | 0,13 |
| **solv** | 0,29 | 0,11 | 0,10 | 0,16 | 0,17 | 0,16 | 0,03 | 0,39 |
| **prem** | 0,17 | 0,21 | 0,22 | 0,10 | 0,10 | 0,00 | 0,03 | 0,15 |
| **act** | 0,01 | 0,29 | 0,30 | 0,13 | 0,12 | 0,01 | 0,04 | 0,06 |
| **num\_all\_new** | 0,14 | 0,09 | 0,10 | 0,04 | 0,04 | 0,01 | 0,06 | 0,14 |
| **num\_all** | 0,15 | 0,16 | 0,16 | 0,09 | 0,09 | 0,00 | 0,05 | 0,14 |
| **prem\_div\_insure\_sum\_new** | 0,09 | 0,04 | 0,04 | 0,02 | 0,02 | 0,17 | 0,08 | 0,12 |
| **ca** | 0,17 | 0,00 | 0,01 | 0,01 | 0,02 | 0,14 | 0,09 | 0,21 |
| **fs6^1\_2** | 0,24 | 0,03 | 0,03 | 0,01 | 0,02 | 0,04 | 0,09 | 0,16 |
| **pr1^1\_2** | 0,06 | 0,28 | 0,28 | 0,35 | 0,35 | 0,02 | 0,23 | 0,09 |
| **pr4^1\_2** | 0,43 | 0,14 | 0,13 | 0,08 | 0,07 | 0,25 | 0,45 | 0,86 |
| **re1^1\_2** | 0,17 | 0,05 | 0,04 | 0,01 | 0,02 | 0,04 | 0,17 | 0,28 |
| **re2^1\_2** | 0,09 | 0,03 | 0,03 | 0,01 | 0,01 | 0,15 | 0,08 | 0,07 |
| **inv^1\_2** | 0,36 | 0,07 | 0,06 | 0,02 | 0,02 | 0,28 | 0,30 | 0,84 |
| **liq1^1\_2** | 0,05 | 0,04 | 0,04 | 0,04 | 0,04 | 0,13 | 0,05 | 0,12 |
| **liq2^1\_2** | 0,60 | 0,00 | 0,00 | 0,13 | 0,14 | 0,08 | 0,14 | 0,38 |
| **delta\_prem^1\_2** | 0,06 | 0,08 | 0,08 | 0,08 | 0,07 | 0,04 | 0,11 | 0,03 |

Табл. 11

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **re3** | **solv** | **prem** | **act** | **num\_all\_new** | **num\_all** | **prem\_div**  **\_insure\_sum\_new** | **ca** |
| **fs3** | 0,03 | 0,29 | 0,17 | 0,01 | 0,14 | 0,15 | 0,09 | 0,17 |
| **pr2** | 0,04 | 0,11 | 0,21 | 0,29 | 0,09 | 0,16 | 0,04 | 0,00 |
| **pr3** | 0,04 | 0,10 | 0,22 | 0,30 | 0,10 | 0,16 | 0,04 | 0,01 |
| **pr5** | 0,00 | 0,16 | 0,10 | 0,13 | 0,04 | 0,09 | 0,02 | 0,01 |
| **pr6** | 0,00 | 0,17 | 0,10 | 0,12 | 0,04 | 0,09 | 0,02 | 0,02 |
| **unpr1** | 0,09 | 0,16 | 0,00 | 0,01 | 0,01 | 0,00 | 0,17 | 0,14 |
| **unpr2** | 0,09 | 0,03 | 0,03 | 0,04 | 0,06 | 0,05 | 0,08 | 0,09 |
| **unpr3** | 0,13 | 0,39 | 0,15 | 0,06 | 0,14 | 0,14 | 0,12 | 0,21 |
| **re3** | 1,00 | 0,25 | 0,05 | 0,04 | 0,05 | 0,06 | 0,12 | 0,09 |
| **solv** | 0,25 | 1,00 | 0,14 | 0,03 | 0,08 | 0,10 | 0,03 | 0,05 |
| **prem** | 0,05 | 0,14 | 1,00 | 0,86 | 0,78 | 0,85 | 0,09 | 0,66 |
| **act** | 0,04 | 0,03 | 0,86 | 1,00 | 0,62 | 0,69 | 0,07 | 0,69 |
| **num\_all\_new** | 0,05 | 0,08 | 0,78 | 0,62 | 1,00 | 0,89 | 0,16 | 0,57 |
| **num\_all** | 0,06 | 0,10 | 0,85 | 0,69 | 0,89 | 1,00 | 0,12 | 0,56 |
| **prem\_div\_insure\_sum\_new** | 0,12 | 0,03 | 0,09 | 0,07 | 0,16 | 0,12 | 1,00 | 0,17 |
| **ca** | 0,09 | 0,05 | 0,66 | 0,69 | 0,57 | 0,56 | 0,17 | 1,00 |
| **fs6^1\_2** | 0,06 | 0,17 | 0,18 | 0,11 | 0,13 | 0,18 | 0,19 | 0,20 |
| **pr1^1\_2** | 0,09 | 0,20 | 0,09 | 0,06 | 0,09 | 0,08 | 0,05 | 0,04 |
| **pr4^1\_2** | 0,13 | 0,37 | 0,15 | 0,04 | 0,14 | 0,13 | 0,13 | 0,20 |
| **re1^1\_2** | 0,48 | 0,15 | 0,15 | 0,16 | 0,12 | 0,09 | 0,26 | 0,34 |
| **re2^1\_2** | 0,03 | 0,14 | 0,09 | 0,03 | 0,09 | 0,11 | 0,10 | 0,07 |
| **inv^1\_2** | 0,14 | 0,35 | 0,16 | 0,09 | 0,16 | 0,16 | 0,03 | 0,22 |
| **liq1^1\_2** | 0,11 | 0,16 | 0,08 | 0,07 | 0,07 | 0,06 | 0,05 | 0,11 |
| **liq2^1\_2** | 0,13 | 0,31 | 0,21 | 0,07 | 0,17 | 0,19 | 0,07 | 0,20 |
| **delta\_prem^1\_2** | 0,03 | 0,18 | 0,16 | 0,10 | 0,16 | 0,13 | 0,10 | 0,07 |

Табл. 12

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **fs6^1\_2** | **pr1^1\_2** | **pr4^1\_2** | **re1^1\_2** | **re2^1\_2** | **inv^1\_2** | **liq1^1\_2** | **liq2^1\_2** |
| **fs3** | 0,24 | 0,06 | 0,43 | 0,17 | 0,09 | 0,36 | 0,05 | 0,60 |
| **pr2** | 0,03 | 0,28 | 0,14 | 0,05 | 0,03 | 0,07 | 0,04 | 0,00 |
| **pr3** | 0,03 | 0,28 | 0,13 | 0,04 | 0,03 | 0,06 | 0,04 | 0,00 |
| **pr5** | 0,01 | 0,35 | 0,08 | 0,01 | 0,01 | 0,02 | 0,04 | 0,13 |
| **pr6** | 0,02 | 0,35 | 0,07 | 0,02 | 0,01 | 0,02 | 0,04 | 0,14 |
| **unpr1** | 0,04 | 0,02 | 0,25 | 0,04 | 0,15 | 0,28 | 0,13 | 0,08 |
| **unpr2** | 0,09 | 0,23 | 0,45 | 0,17 | 0,08 | 0,30 | 0,05 | 0,14 |
| **unpr3** | 0,16 | 0,09 | 0,86 | 0,28 | 0,07 | 0,84 | 0,12 | 0,38 |
| **re3** | 0,06 | 0,09 | 0,13 | 0,48 | 0,03 | 0,14 | 0,11 | 0,13 |
| **solv** | 0,17 | 0,20 | 0,37 | 0,15 | 0,14 | 0,35 | 0,16 | 0,31 |
| **prem** | 0,18 | 0,09 | 0,15 | 0,15 | 0,09 | 0,16 | 0,08 | 0,21 |
| **act** | 0,11 | 0,06 | 0,04 | 0,16 | 0,03 | 0,09 | 0,07 | 0,07 |
| **num\_all**  **\_new** | 0,13 | 0,09 | 0,14 | 0,12 | 0,09 | 0,16 | 0,07 | 0,17 |
| **num\_all** | 0,18 | 0,08 | 0,13 | 0,09 | 0,11 | 0,16 | 0,06 | 0,19 |
| **Prem\_div**  **\_insure\_**  **sum\_new** | 0,19 | 0,05 | 0,13 | 0,26 | 0,10 | 0,03 | 0,05 | 0,07 |
| **ca** | 0,20 | 0,04 | 0,20 | 0,34 | 0,07 | 0,22 | 0,11 | 0,20 |
| **fs6^1\_2** | 1,00 | 0,11 | 0,23 | 0,22 | 0,05 | 0,32 | 0,16 | 0,23 |
| **pr1^1\_2** | 0,11 | 1,00 | 0,03 | 0,10 | 0,02 | 0,11 | 0,11 | 0,12 |
| **pr4^1\_2** | 0,23 | 0,03 | 1,00 | 0,27 | 0,07 | 0,81 | 0,11 | 0,39 |
| **re1^1\_2** | 0,22 | 0,10 | 0,27 | 1,00 | 0,30 | 0,32 | 0,19 | 0,21 |
| **re2^1\_2** | 0,05 | 0,02 | 0,07 | 0,30 | 1,00 | 0,07 | 0,02 | 0,01 |
| **inv^1\_2** | 0,32 | 0,11 | 0,81 | 0,32 | 0,07 | 1,00 | 0,20 | 0,35 |
| **liq1^1\_2** | 0,16 | 0,11 | 0,11 | 0,19 | 0,02 | 0,20 | 1,00 | 0,10 |
| **liq2^1\_2** | 0,23 | 0,12 | 0,39 | 0,21 | 0,01 | 0,35 | 0,10 | 1,00 |
| **delta\_**  **prem^1\_2** | 0,25 | 0,24 | 0,01 | 0,11 | 0,09 | 0,03 | 0,09 | 0,04 |

В выборке сильно коррелированными переменными являются: pr2, pr3, pr5, pr6, unpr3, pr4^1\_2, inv^1\_2, prem, act, num\_all\_new, num\_all и act. Чтобы не дублировать похожие переменные и не создавать мультиколлинеарность, в модели из выборки удалены переменные с наименьшей объясняющей силой, а именно с меньшим коэффициентом Джини. В таблице ниже (Табл. 13) приведены попарно сильно коррелируемые переменные (корреляция больше 0,75), причем значение коэффициента Джини второго параметра меньше, чем первого. Следовательно, 8 различных параметров, находящихся во второй колонке, должны быть удалены из выборки, но если присмотреться повнимательнее к таблице, то можно заметить, что в выборке можно оставить переменную количества активов, так как она сильно коррелирует только с фактором размера премий, который в свою очередь удаляется из выборки за счет переменной количества договоров.

Табл. 13

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **param1** | **param2** | **Corr** | **gini1** | **gini2** |
| pr2 | pr3 | 0,980 | 0,326 | 0,314 |
| pr2 | pr5 | 0,839 | 0,326 | 0,310 |
| pr2 | pr6 | 0,853 | 0,326 | 0,322 |
| pr3 | pr5 | 0,860 | 0,314 | 0,310 |
| pr6 | pr3 | 0,840 | 0,322 | 0,314 |
| pr6 | pr5 | 0,981 | 0,322 | 0,310 |
| prem | act | 0,863 | 0,215 | 0,199 |
| num\_all\_new | prem | 0,784 | 0,280 | 0,215 |
| num\_all | prem | 0,848 | 0,334 | 0,215 |
| num\_all | num\_all\_new | 0,885 | 0,334 | 0,280 |
| pr4^1\_2 | unpr3 | 0,855 | 0,317 | 0,109 |
| pr4^1\_2 | inv^1\_2 | 0,814 | 0,317 | 0,132 |
| inv^1\_2 | unpr3 | 0,837 | 0,132 | 0,109 |

## **4.4. Дамми-показатели**

Для лучшей разделительной способности модели и отслеживания кризисных годов или, наоборот, поднятия экономики, добавим категориальную переменную года. А именно с 2012 по 2016 год, не добавляя бинарную переменную, отвечающую за 2017 года, так как в обучающей выборке данный столбец будет полностью состоять из нулей, и модель на нем обучаться не будет.

2012 и 2013 гг. были годами подъема экономики, а точнее выхода на докризисный уровень 2007 года, с 2014 года начался еще один кризис в экономике России.

Так же в модели используются переменные вида деятельности страховой компании, перечисленные в таблице ниже (Табл. 14). Информация о страховой деятельности компании получена из xml-файлов годовой финансовой отчетности из Формы 0 с основным описанием компании [37] и с сайта ЦБ с информацией о действующих лицензиях страховых компаний на данный момент.

Табл. 14

|  |  |
| --- | --- |
| **Обозначение** | **Вид деятельности** |
| life | Страхование жизни |
| medical | Обязательное медицинское страхование |
| medical\_voluntary | Добровольное медицинское страхование |
| avto | Обязательное страхование гражданской ответственности владельцев транспортных средств |
| liability | Страхование гражданской ответственности владельцев воздушного, водного, железнодорожного транспорта и организаций, эксплуатирующих опасные объекты |
| exceptlife\_other | Прочие виды страхования иного, чем страхование жизни |
| reinsurance\_in | Перестрахование принятое |
| reinsurance\_out | Перестрахование переданное |
| mutual | Взаимное страхование |

Ожидается отрицательное воздействие переменных вида деятельности страхования жизни и медицинского, и отрицательное воздействие в эти годы страхования ОСАГО в виду начавшегося кризиса в этой отрасли на вероятность дефолта страховщика.

(reg\_mos) Из адреса региона регистрации страховой компании была получена региональная переменная, равная «1», если регион регистрации Москва, и «0» - иначе.

Макропеременные, такие как ВВП, инфляция, и другие, в работе не использовались, так как суть работы – это спрогнозировать дефолты за 2019 год по годовой финансовой отчетности за 2018 год, а на конец 2019 года информация по данным макропеременным на дату написания дипломной работы отсутствует, в то время как делать экспертные оценки этих факторов не входит в мою область компетенции.

# **5. Моделирование**

Получившаяся итоговая база данных содержит 2025 наблюдений с 2012 по 2017 год и 33 различные объясняющие переменные. Разделим выборку на обучающую и тестовую, где в обучающую входят все наблюдения, кроме 2017 года. (Табл. )

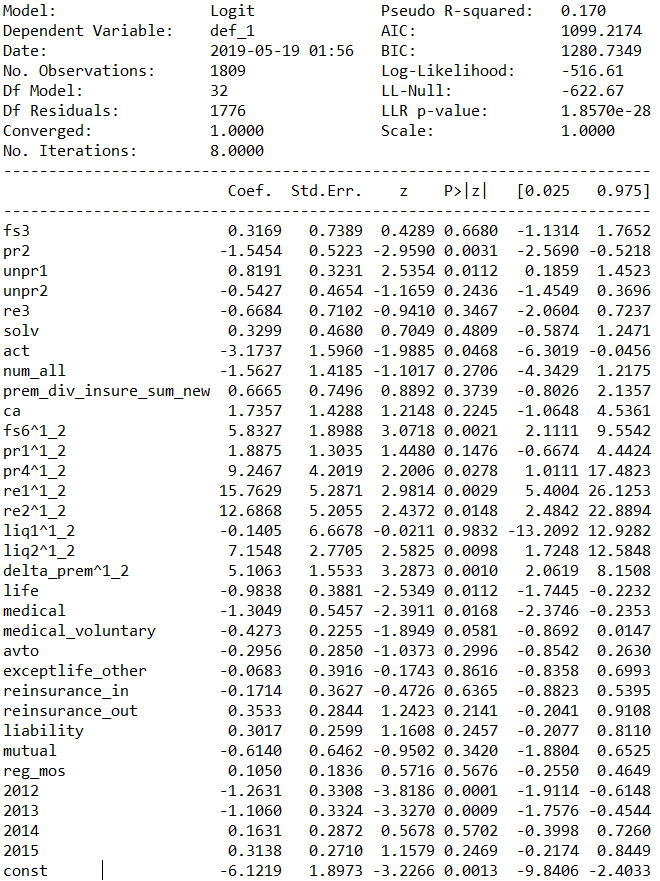
Табл. 15

|  |  |  |
| --- | --- | --- |
|  | **Количество наблюдений** | **Доля дефолтов, %** |
| Вся выборка  (2012-2017 гг.) | 2025 | 11.16% |
| Обучающая выборка (2012-2016 гг.) | 1809 | 10.89% |
| Тестовая выборка (2017 г.) | 216 | 13.42% |

## **5.1. Логит-модель от всех переменных (Модель 0)**

Построим логит-модель на обучающей выборке от всех переменных, при чем 2016 год возьмем за базовый и добавим в модель константу. Значение коэффициента Джини на обучающей выборке равно 58,7%, а на тестовой 56,7%, что является достаточно хорошим результатом. Но переменные fs3, unpr2, re3, solv, num\_all, prem\_div\_insure\_sum\_new, ca, pr1^1\_2, liq1^1\_2, avto, exceptlife\_other, reinsurance\_out, reinsurance\_in, liability, mutual, reg\_mos, 2014 и 2015 года не значимы на уровне значимости 10%. А по фактору medical\_voluntary ноль попадает в доверительный интервал, что так же может говорить о его незначимости. (Таблица 16)

Табл. 16



По графику ROC-кривой тестовой выборки можно заметить, что модель ошибочно проставила наибольшую вероятность банкротства компаниям, у которых не отняли лицензию в течение следующего года. При чем в дальнейшем при уменьшении прогноза модель достаточно хорошо угадывает дефолты, а именно соотносит спрогнозированную вероятность дефолта с текущим финансовым состоянием компании. (Рис. )

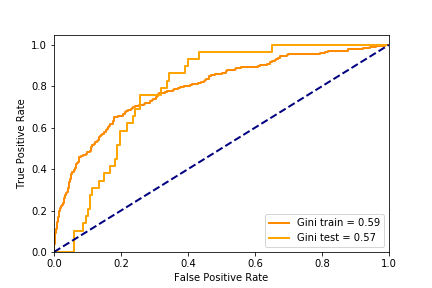
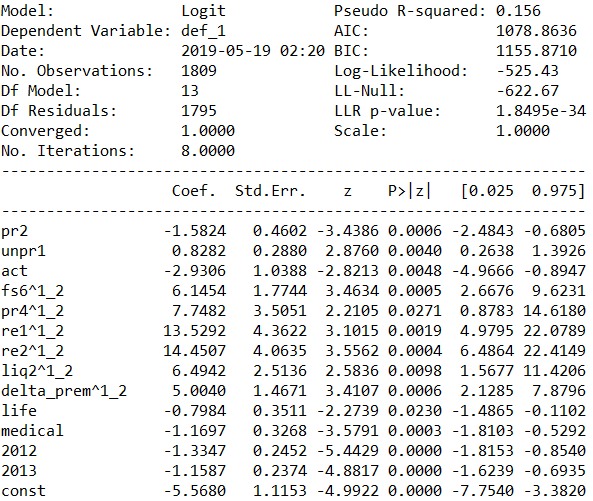


Рис. 17

## **5.2. Логит-модель от всех значимых переменных (Модель 1)**

Удалим все незначащие переменные и построим Логит-модель на оставшихся переменных. Джини полученной обучающей выборки равен 55,6%, а тестовой – 57,6%, то есть модель стала чуть хуже распознавать дефолты на обучающей выборке, но зато на тестовой выборке Джини значительно вырос, а также все переменные значимы на уровне значимости 10%, и направление влияния всех переменных совпадает ожидаемым.

Табл. 17



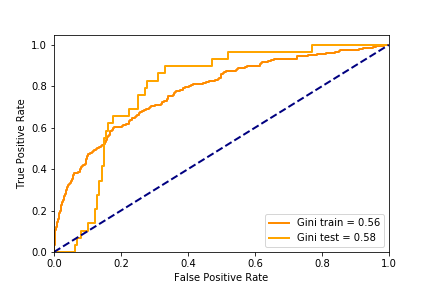


Рис. 18

Результат графика (Рис. 18) похож на предыдущий (Рис 17) и, следовательно, выводы те же.

Все vif в новой модели по всем переменным меньше 3, следовательно, в модели отсутствует мультиколлинеарность. (Табл. 18)

Табл. 18

|  |  |
| --- | --- |
| **Param** | **VIF** |
| **pr2** | 1,23 |
| **unpr1** | 1,20 |
| **act** | 1,27 |
| **fs6^1\_2** | 1,28 |
| **pr4^1\_2** | 1,65 |
| **re1^1\_2** | 1,44 |
| **re2^1\_2** | 1,18 |
| **liq2^1\_2** | 1,48 |
| **delta\_prem^1\_2** | 1,12 |
| **life** | 1,21 |
| **medical** | 2,27 |
| **2012** | 1,17 |
| **2013** | 1,14 |

## **5.3. RFECV метод для Логит-модели (Модель 2.1 и Модель 2.2)**

Вероятно, данный набор переменных является не окончательным, так как возможно мы выбросили нужные факторы или наоборот оставили лишние. Поэтому попробуем применить метод RFECV ранжирования переменных с рекурсивным удалением факторов и перекрестным проверкой выбора лучшего количества функций. На вход данному методу подается наша обучающая выборка, задается метод оценки, как roc\_auc, и логит модель. Также указывается на сколько частей делиться выборка для кросс-валидации.

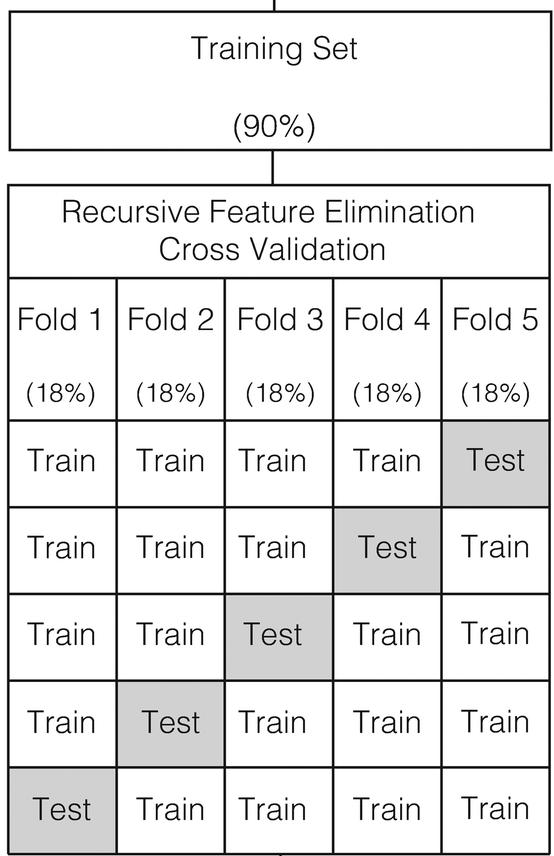
Метод RFECV используется для определения того, какие факторы важны для прогнозирования вероятности дефолта, и чтобы избежать переобучения за счет сокращения количества переменных, используемых для обучения. По своей сути база данных, на которой обучается модель, должна включать репрезентативный набор факторов. Тем не менее, изначально неясно, какие переменные важны для прогнозирования, и необходимо предпринять шаги для уменьшения чрезмерной подгонки. Использование RFECV для выявления и удаления признаков, имеющих малую значимость для прогнозирования, позволяет «обрезать» учебную базу.

Рис. 19

Для каждого прогона модели обучающий набор делится на 5 случайно разделенных подмножеств (cross-validation folds). Каждое кросс-валидационное подмножество состоит из 20% обучающего набора данных или 18% от общей базы данных. Четыре обучающих подмножества объединяются для формирования временной базы данных для перекрестной проверки, на которой обучается логит-модель. (Рис. 19) Производительность модели затем рекурсивно оценивается путем систематического удаления признаков с наименьшей значимостью из обучающего набора, повторного обучения модели, а затем проверки прогнозирующей эффективности модели с использованием оставшегося 5-го тестового подмножества. Для каждого шага рекурсии внутри итерации исключается элемент с наименьшей важностью, а логит-модель проходит повторное обучение и повторную проверку, пока не останется один элемент.

Кросс-валидационная оценка моделей при различном количестве переменных от 1 до 34 при разбиении обучающей выборки на 5 частей представлена ниже. (Рис. 20)

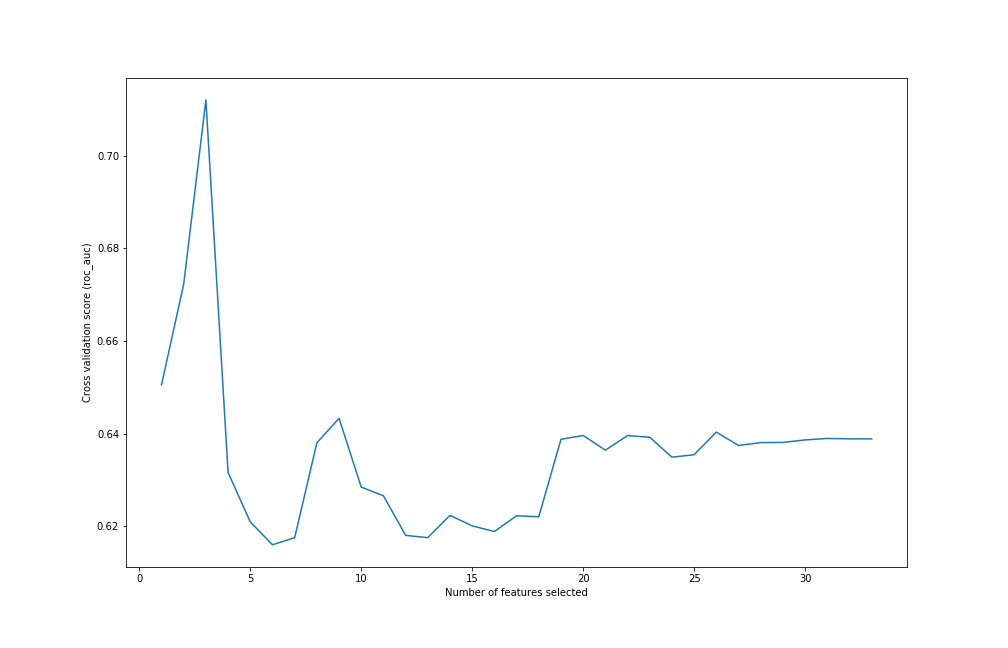
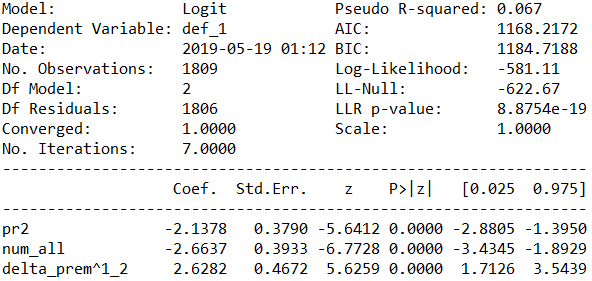


Рис. 20

Оптимальным количеством переменных метод признал три переменные. Описание полученной логит-модели представлено ниже (Табл. 19). Значение коэффициента Джини на обучающей выборке равно 39,1%, а на тестовой – 57,2%. Можно сделать вывод, что на обучающей выборке модель явно плохо работает. (Модель 2.1)

Табл. 19



Если обучающую выборку делить только на 2 части, то оптимальным числом факторов является 4. Кросс-валидационная оценка модели и описательная статистика представлены ниже (Рис. 21 и Табл. 20). (Модель 2.2)

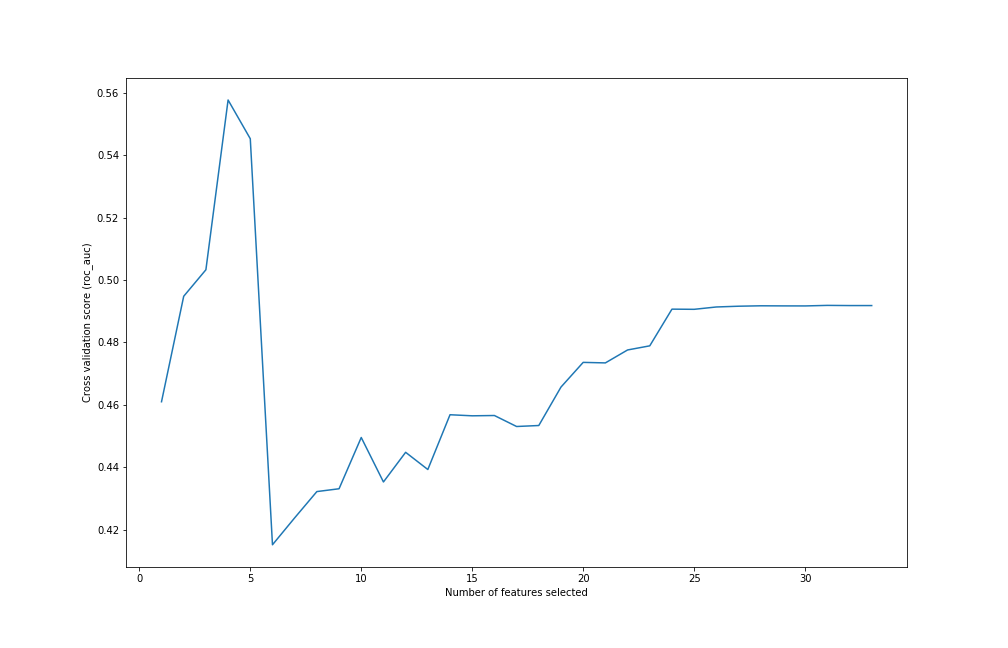
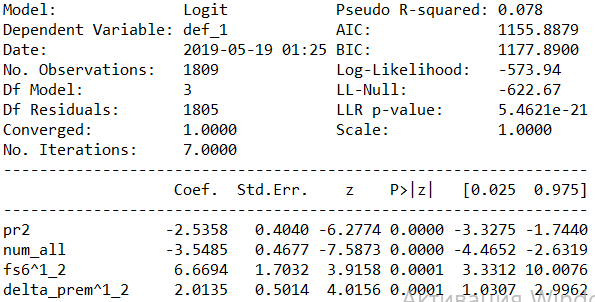


Рис. 21

Табл. 20



Причем все факторы являются значимыми на любом уровне значимости, коэффициент Джини на обучающей выборке равен всего лишь 41,1%, а на тестовой достигает 60,3 балла.

## **5.4. Перебор логит-моделей (Модели 3.1, 3.2 и 3.3)**

При большом количестве факторов в моделях показатели в большинстве своем не значимы, но при уменьшении их количества показатель Gini зачастую тоже снижается, поэтому необходимо подобрать их оптимальное число. Если перебирать все комбинации, то количество комбинаций при выборе 6 показателей из 34 превышает 1 000 000, и стандартному офисному компьютеру потребуется около суток, чтобы провести вычисление и оценку. Данное быстродействие не может нас устраивать. Поэтому был предложен упрощенный способ для выбора оптимальной модели.

Для поиска наилучшей модели построим все комбинации объясняющих переменных, состоящих из 5, 6 или 7-ми финансовых факторов из всех отобранных 18 и добавим к этим наборам факторы страховой деятельности, такие как страхование жизни и медицинское, а также добавим два докризисных года (2012 и 2013) и константу. В итоге получится 58956 моделей, по которым рассчитаем коэффициент Джини на обучающей выборке и тестовой. При чем у 12 моделей Джини на тесте равен 70-71%, а у 9680 моделей – больше 60%. На обучающей выборке результаты не такие прекрасные, так только у 27 моделей Джини доходит до 55%.

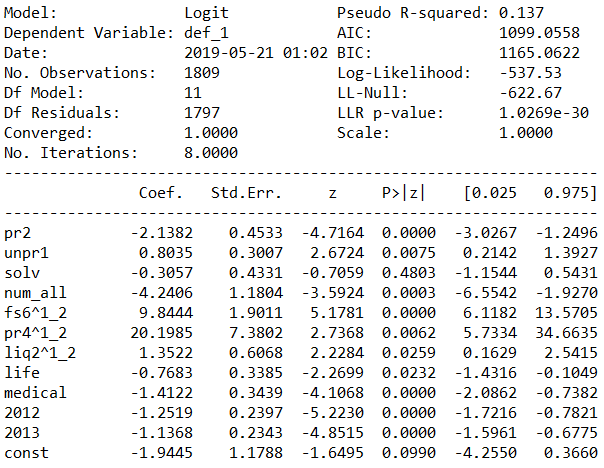
Выпишем 5 лучших моделей, отобранных по наибольшему значению коэффициента Джини на тестовой выборке. (Табл. 21)

Табл. 21

|  |  |  |  |
| --- | --- | --- | --- |
| **Параметры** | **Кол-во** | **gini\_tr** | **gini\_ts** |
| ;pr2;unpr1;solv;num\_all;fs6^1\_2;pr4^1\_2;liq2^1\_2;life;  medical;2012;2013;const; | 12 | 52,7% | 71,0% |
| ;pr2;unpr1;num\_all;fs6^1\_2;pr4^1\_2;re1^1\_2;liq2^1\_2;life;  medical;2012;2013;const; | 12 | 53,7% | 70,9% |
| ;pr2;unpr1;num\_all;fs6^1\_2;pr4^1\_2;liq1^1\_2;liq2^1\_2;life;  medical;2012;2013;const; | 12 | 52,8% | 70,6% |
| ;pr2;unpr1;re3;num\_all;fs6^1\_2;pr4^1\_2;liq2^1\_2;life;  medical;2012;2013;const; | 12 | 52,7% | 70,6% |
| ;pr2;unpr1;num\_all;fs6^1\_2;pr1^1\_2;pr4^1\_2;liq2^1\_2;life;  medical;2012;2013;const; | 12 | 53,8% | 70,5% |

Рассмотрим статистику по первой модели (Табл. 22) (Модель 3.1)

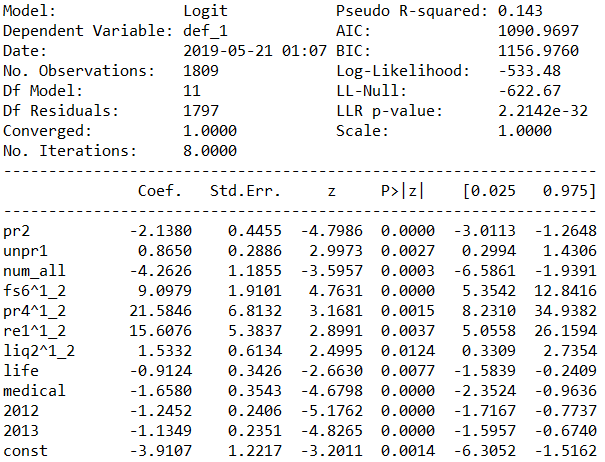
Табл. 22



У данной модели показатель платежеспособности не значим, но при его исключении новая модель уже явно в топе не находится (так как мы построили прогнозы и для всех моделей от различных комбинаций 11 переменных), а значит и коэффициент Джини у нее будет заметно меньше.

Перейдем к следующей модели. (Модель 3.2)

Табл. 23



У данной модели все переменные значимы на уровне 5%, и знаки коэффициентов объясняющих переменных совпадают с нашими экспертными ожиданиями. (Табл. 23)

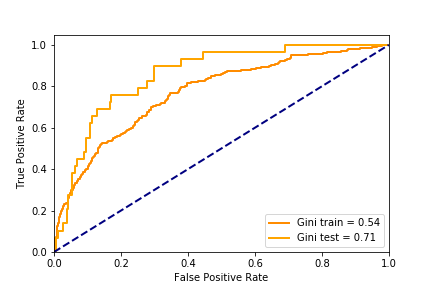


Рис. 22

Рассматривая ROC-кривую, можно заметить, что она ни разу не пересекает диагональ прямоугольника, а именно ROC-кривую модели случайного угадывания, а значит, что спрогнозированная вероятность дефолта коррелирует и сонаправлена с объясняемой переменной на всех участках. (Рис. 22)

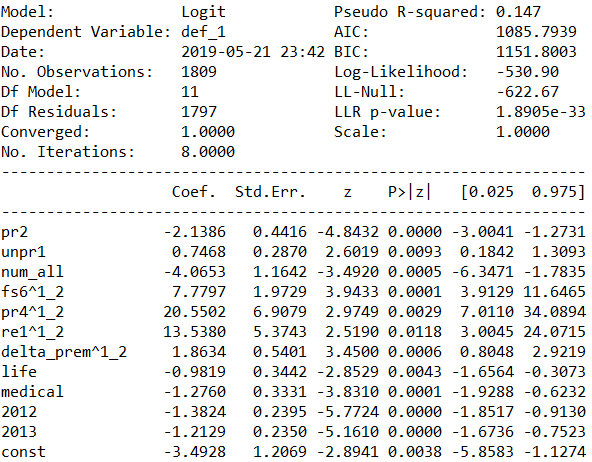
Отберем модели по ограничению вида: Джини на тесте больше 56%, а на обучающей больше 55%. Тогда получим 15 моделей, представленных ниже. (Табл. 24)

Табл. 24

|  |  |  |  |
| --- | --- | --- | --- |
| **param** | **num** | **gini\_tr** | **gini\_ts** |
| ;pr2;unpr1;num\_all;fs6^1\_2;pr4^1\_2;re1^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 12 | 55,4% | 66,6% |
| ;pr2;unpr2;num\_all;fs6^1\_2;pr4^1\_2;re1^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 12 | 55,1% | 66,5% |
| ;pr2;num\_all;fs6^1\_2;pr1^1\_2;pr4^1\_2;re1^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 12 | 55,7% | 66,0% |
| ;pr2;act;num\_all;fs6^1\_2;pr4^1\_2;re1^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 12 | 55,1% | 65,7% |
| ;pr2;num\_all;fs6^1\_2;pr4^1\_2;re1^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 11 | 55,1% | 65,6% |
| ;pr2;num\_all;ca;fs6^1\_2;pr4^1\_2;re1^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 12 | 55,0% | 65,6% |
| ;pr2;num\_all;prem\_div\_insure\_sum\_new;fs6^1\_2;pr4^1\_2;re1^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 12 | 55,2% | 65,4% |
| ;pr2;num\_all;fs6^1\_2;pr4^1\_2;re1^1\_2;liq1^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 12 | 55,0% | 65,3% |
| ;pr2;unpr1;num\_all;fs6^1\_2;pr1^1\_2;pr4^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 12 | 55,1% | 65,2% |
| ;pr2;num\_all;prem\_div\_insure\_sum\_new;fs6^1\_2;pr1^1\_2;pr4^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 12 | 55,3% | 65,2% |
| ;pr2;solv;num\_all;fs6^1\_2;pr4^1\_2;re1^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 12 | 55,1% | 64,4% |
| ;fs3;pr2;num\_all;fs6^1\_2;pr4^1\_2;re1^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 12 | 55,5% | 64,3% |
| ;fs3;pr2;num\_all;fs6^1\_2;pr1^1\_2;pr4^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 12 | 55,1% | 64,2% |
| ;pr2;solv;num\_all;fs6^1\_2;pr1^1\_2;re1^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 12 | 55,0% | 62,1% |
| ;fs3;pr2;num\_all;fs6^1\_2;pr1^1\_2;re1^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const; | 12 | 55,3% | 61,9% |

Рассмотрим первую модель с gini\_train = 55.4% и gini\_test = 66.6%. (Модель 3.3) (Табл. 25) (Рис. 23)

Табл. 25



Она обладает достаточно хорошей предсказательной способностью, псевдо- R-squared вырос относительно двух предыдущих моделей, коэффициенты Джини являются достаточно большими и все переменные значимы и коэффициенты перед ними имеют нужный знак.

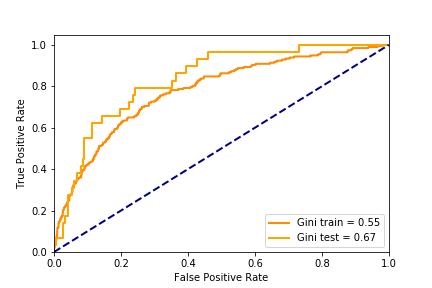


Рис. 23

При переборе моделей с 13 и 14 переменными включая константу было построено и обучено 92378 логит-моделей. При чем Джини на тесте доходит до 72%, а на обучающей выборке до 56,4%. Но почти все эти модели имеют pseudo-R-squared примерно равным 0,144, и как минимум одну незначащую переменную. А такие модели нам не подходят.

## **5.5. Модель с машинным обучением (Модель 4)**

Методология дерева решений (Decision tree) - это широко используемый метод интеллектуального анализа данных для создания систем классификации на основе нескольких ковариант или для разработки алгоритмов прогнозирования целевой переменной. Этот метод классифицирует выборку на ветко-образные сегменты, которые создают перевернутое дерево с корневым узлом, внутренними узлами и конечными узлами-листьями. Алгоритм является непараметрическим и может эффективно работать с большими сложными наборами данных без наложения сложной параметрической структуры. Хотя размер нашей выборки недостаточно велик, данные для исследования перед применением данного алгоритма все равно разделим на наборы данных для обучения и проверки. Обучающий набор данных используем для построения модели дерева решений и тестовый набор - для выбора подходящего размера дерева, необходимого для достижения оптимальной конечной модели и лучших результатов.

An external file that holds a picture, illustration, etc.
Object name is sap-27-02-130-g002.jpgНа рисунке слева показана простая модель дерева решений, которая включает в себя одну двоичную целевую переменную Y (0 или 1) и две непрерывные переменные, x1 и x2, которые находятся в диапазоне от 0 до 1. (Рис. 24) Основными компонентами модели дерева решений являются узлы и ветви и наиболее важными шагами в построении модели являются разбиение, остановка и обрезка.

Рис. 24

После перебора достаточно большого кол-ва моделей решено было остановиться на Дереве решений с параметрами конфигурации:

* максимальная глубина – 20;
* минимальное кол-во наблюдений в листьях – 1;
* минимальное кол-во наблюдений в узлах – 4.

Так как количество узлов в дереве равно 310, то саму схему дерева выписывать не будем. Джини по такому дереву на обучающей выборке 92,46%, а на тестовой – 17,48%. То есть данный алгоритм очень хорошо распознает и классифицирует обучающую выборку, но совсем не пригоден на тестовой. (Табл. 26) Табл. 26

|  |  |  |
| --- | --- | --- |
|  | **0\_fact** | **1\_fact** |
| **0\_pred** | 181 | 23 |
| **1\_pred** | 6 | 6 |

При чем дефолты правильно угадывает только в 20% случаях, а не дефолт в 96%, вследствие, большого кол-ва не дефолтных компаний в выборке.

Данная модель обладает достаточно плохой предсказательной способностью вероятно из-за того, что в одной и той же ветке дерева могут встретиться как минимум два ограничения в узлах на одну и туже финансовую переменную, то есть, к примеру, модель будет считать наблюдением не дефолтным при значении переменной от 0 до 0,3, затем от 0,3 до 0,6 – дефолтным, и оставшийся отрезок – не дефолтным. Хотя это идет в разрез нашим предположениям о постоянном направлении влиянии финансовых факторов и свидетельствует о переобучении модели на обучающей выборке. При чем при попытке выставлении более строгого ограничения на глубину дерева его предсказательная способность падает не только на обучающей выборке, но и на тестовой.

## **5.6. Сравнение моделей**

По результатам сравнения roc-кривых, коэффициентов Джини, псевдо R-квадратов на обучающей и проверочной выборке решено было лучшими моделями признать Модель 3.2 и Модель 3.3 со всеми значащими переменными. И продолжить дальнейший выбор конечной модели только среди них.

Для итогового построения прогноза необходимо определить границу вероятности дефолта, больше которой наблюдение считается дефолтным, и меньше - наоборот. Для оценки качества новых прогнозов будем опять пользоваться метрикой Джини. Для этого на обучающей выборке циклично границу вероятности дефолта (P\*) будем приравнивать ко всем возможным вероятностям дефолта, присвоенным наблюдениям из обучающей выборки, и рассчитывать новый получившийся коэффициент Джини при заданной границе.

Тогда Модели 3.2 соответствует график (Рис. 25), представленный ниже, где по оси X брались границы вероятности дефолта (P\*), а по оси Y – новый получившийся Джини обучающей выборки при предсказаниях, равных только «0» или «1».

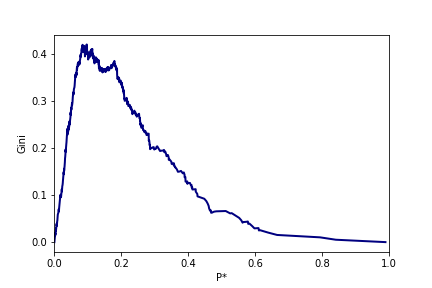


Рис. 25

В итоге для данной модели на обучающей выборке наилучшей разделительной способностью обладает граница P\* = 0.0977.

Тогда на обучающей выборке модель 3.2 правильно распознает 65,4% недефолтных компаний, и 76,1% дефолтных. (Табл. 27)

Табл. 27

|  |  |  |
| --- | --- | --- |
|  | **0\_fact** | **1\_fact** |
| **0\_pred** | 1054 | 47 |
| **1\_pred** | 558 | 150 |

А на тестовой выборке модель 3.2 правильно распознает 67,9% недефолтных компаний, и 89,7% дефолтных. (Табл. 28) При чем ошибками первого рода являются 60 компаний, а второго рода – только 3.

Табл. 28

|  |  |  |
| --- | --- | --- |
|  | **0\_fact** | **1\_fact** |
| **0\_pred** | 127 | 3 |
| **1\_pred** | 60 | 26 |

Коэффициенты Джини на обоих выборках с классифицированными вероятностями дефолтов равны 41,53% и 57,57% соответственно.

Рассмотрим Модель 3.3 и для нее наилучшей границей P\* является 10,77%. (Рис. 26)

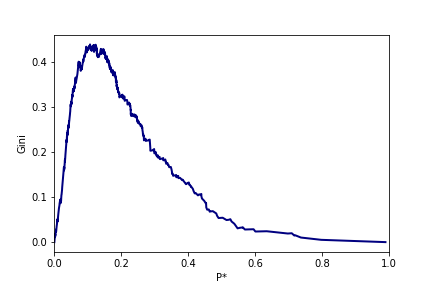
****

Рис. 26

Тогда на обучающей выборке модель 3.2 правильно распознает 68,2% недефолтных компаний, и 75,1% дефолтных. (Табл. 29)

Табл. 29

|  |  |  |
| --- | --- | --- |
|  | **0\_fact** | **1\_fact** |
| **0\_pred** | 1100 | 49 |
| **1\_pred** | 512 | 148 |

А на тестовой выборке модель 3.2 правильно распознает 68,4% недефолтных компаний, и 79,3% дефолтных. (Табл. 30) При чем ошибками первого рода являются 59 компаний, а второго рода – только 6.

Табл. 30

|  |  |  |
| --- | --- | --- |
|  | **0\_fact** | **1\_fact** |
| **0\_pred** | 128 | 6 |
| **1\_pred** | 59 | 23 |

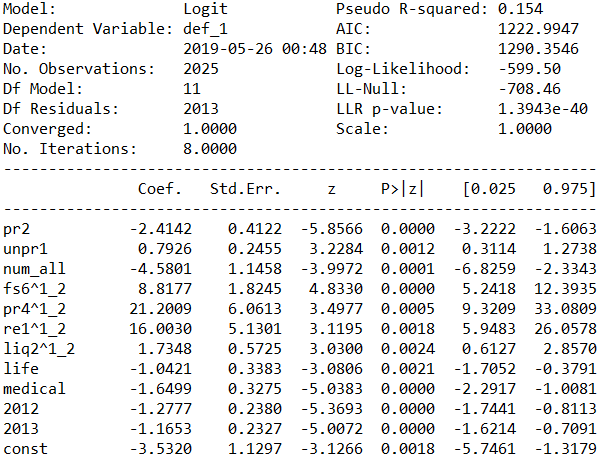
Коэффициенты Джини на обоих выборках с классифицированными вероятностями дефолтов равны 43,37% и 47,76% соответственно.

Принимая во внимание, что результатами данной модели будут пользоваться потребители страховых услуг, то для них лучше, если они с большей точностью смогут определять будущие дефолтные компании, и им не так критично, если часть «хороших» компаний тоже попадет в так называемый «черный» список, так как все равно останется еще больше 128 компаний, к которым они уверенно смогут обраться. Следовательно, так как по результатам модели 3.2 ошибки второго рода на обучающей и тестовой выборке меньше, то данная модель была признана наилучшей.

## **5.7. Описание итоговой модели**

На этапе описания модели необходимо обучить Модель 3.2 всей выборке. Статистика по данной модели представлена в таблице 31.

Табл. 31



Псевдо R-квадрат данной модели равен 0.154, а коэффициент Джини 55,98%, можно заметить, что данные показатели выросли относительно показателей модели, обученной только на обучающей выборке. Все переменные и константа остались значимы на уровне значимости 1%, и в доверительный интервал значений коэффициентов 0 тоже не попадает.

Текущая конечная модель состоит из 7 финансовых переменных и 4-ех дамми-переменных. При чем показатель рентабельности собственного капитала, как и количество действовавших договоров страхования на конец отчетного периода воздействуют отрицательно на вероятность наступления дефолта. А показатель убыточности, равный отношению расходов по страховой деятельности к страховым премиям, положительно влияет на вероятность банкротства страховщика. Между показателями соотношения суммы страховых премий и страховых резервов, чистой рентабельностью, долей перестраховщиков в страховых резервах, долей наиболее ликвидных активов в общем объеме активов и вероятностью дефолта существует нелинейная связь, то есть вероятность дефолта растет при очень маленьких или очень больших значениях данных показателей. Так же если компания занимается страхованием жизни или обязательным медицинским страхованием, то это благотворно влияет на ее финансовое состоянии и стабильность. А 2012 и 2013 являются годами подъема экономики в России, относительно спада 2014-2016 годов, и, следовательно, в те года рынок страхования чувствовал себя лучше, чем в последующий период.

При оптимальной границе вероятности дефолтов равной 8,92%, матрица попаданий в нужные значения «дефолта» и «не дефолта», ошибок первого и второго рода на всей выборке принимает следующий вид (Табл.32):

Табл. 32

|  |  |  |
| --- | --- | --- |
|  | **0\_fact** | **1\_fact** |
| **0\_pred** | 1001 | 39 |
| **1\_pred** | 611 | 158 |

Исходя из того, что значение VIF по всем переменным в модели не превышает 3, то в данной модели отсутствует мультиколинеарность. (Табл. 33)

Табл. 33

|  |  |
| --- | --- |
| **Param** | **VIF** |
| pr2 | 1,13 |
| unpr1 | 1,12 |
| num\_all | 1,15 |
| fs6^1\_2 | 1,15 |
| pr4^1\_2 | 1,70 |
| re1^1\_2 | 1,28 |
| liq2^1\_2 | 1,56 |
| life | 1,14 |
| medical | 2,34 |
| 2012 | 1,14 |
| 2013 | 1,11 |

# **6. Прогноз на 2019 год**

Так как ЦБ в марте 2019 года перестал поддерживать сайт личного кабинета страховщика с опубликованными данными бухгалтерской отчетности страховых компаний, не успев при этом, выложить годовую финансовую отчетность за 2018 год, и на остальных ресурсах эта информация тоже отсутствует, то в данной дипломной работе пока не удастся спрогнозировать дефолты страховых за 2019 год.

# **7. Заключение**

В данной работе было рассмотрено огромное число моделей вероятности дефолта страховых компаний и экспертно выбрана наилучшая, которая, благодаря тщательному отбору, заполнению пропусков, нормализации и трансформации объясняющих финансовых переменных, продемонстрировала убедительную предсказательную силу при тестировании вне обучающей выборки и на всей выборке: так данная модель на всей выборке предсказывает более 80% дефолтов страховщиков, а на тестовой – 90%.

Так же в результате однофакторного анализа переменных была обнаружена и доказана квадратическая зависимость вероятности дефолта страховщиков от ряда относительных финансовых показателей:

* соотношение суммы страховых премий и страховых резервов;
* рентабельность страховой и финансово-хозяйственной деятельности;
* чистая рентабельность страховой деятельности;
* доля перестраховщиков в страховых резервах (кроме жизни);
* доля страховых премий, переданных в перестрахование в анализируемом периоде;
* уровень покрытия инвестиционными активами страховых резервов-нетто (кроме страхования жизни);
* текущая ликвидность;
* доля наиболее ликвидных активов в общем объеме активов;
* изменение совокупного объема премии.

За время проведения исследования была выполнена огромная работа по написанию программного обеспечения, на языке программирования Python, для автоматизации конвертации форматов из xml в csv и объединения годовых финансовых отчетностей страховых компаний за различные даты, вычисления относительных финансовых факторов, расчета коэффициента Джини при проведении однофакторного анализа с автоматической записью полученных значений в таблицу, построения графиков roc-кривых, цикличное построение и обучение более 100тыс моделей на различных комбинациях объясняющих факторах с вычислением Джини на тестовой и обучающей выборке и перебора всех возможных разграничивающих значений вероятности дефолта для выбора оптимального значения. Код данной программы представлен в приложении.

# **8. Литература**

1. Закона РФ "Об организации страхового дела в Российской Федерации" от 27.11.1992 N 4015-1 (ред. от 31.12.2017) Доступ из справочно-правовой системы «Консультант Плюс».
2. Указание Банка России "О порядке расчета страховой организацией нормативного соотношения собственных средств (капитала) и принятых обязательств" от 28.07.2015 N 3743-У (ред. от 09.01.2018) (Зарегистрировано в Минюсте России 09.09.2015 N 38865)
3. Приказ Минфина РФ "Об утверждении Положения о порядке расчета страховщиками нормативного соотношения активов и принятых ими страховых обязательств" от 02.11.2001 N 90н (ред. от 08.02.2012)
4. Altman, E.I., 1968. Financial ratios: discriminant analysis and the prediction of corporate bankruptcy. J. Finance 23, 589–609.
5. Sinkey, J.F., 1975. A multivariate analysis of the characteristics of problem banks. J. Finance 30, 21–36.
6. Martin, D., 1977. Early warning of bank failure: a logit regression approach. J. Bank. Finance 1 (3), 249–276.
7. Ohlson, J.A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. J. Account. Res. 18, 109–131.
8. Boyacioglu, M.A., Kara, Y., Baykan, O.K., 2009. Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: a comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. Expert Syst. Appl. 36, 3355–3366.
9. А.М. Карминский, А.В. Костров, Т.Н. Мурзенков. Моделирование вероятности дефолта российских банков с использованием эконометрических методов // М. : Изд. дом Высшей школы экономики, 2012. – 64 с.
10. А.М. Карминский, А.В. Костров. Моделирование вероятности дефолта российских банков: расширенные возможности // Журнал Новой экономической ассоциации. – 2013. - № 1 (17).  С. 64–86.
11. Alexander Karminsky; Alexander Kostrov, 2017, The back side of banking in Russia: forecasting bank failures with negative capital, International Journal of Computational Economics and Econometrics (IJCEE), Vol. 7, No. ½.
12. BarNiv, Ran and Robert A. Hershbarger, 1990, Classifying Financial Distress in the Life Insurance Industry, Journal of Risk and Insurance, 57: 1 10- 136.
13. Chen R., Wong K.A., 2004, The Determinants of Financial Health of Asian Insurance Companies, The Journal of Risk and Insurance, Vol. 71, No. 3 (Sep., 2004), pp. 469-499
14. Le, H. H. & Viviani, J.-L., 2018, Predicting bank failure: An improvement by implementing machine-learning approach on classical ﬁnancial ratios, Research in International Business and Finance. Vol. 44, pp. 16-25.
15. Обзор ключевых показателей страховщиков по итогам 2017 года // Центральный банк Российской Федерации. *https://www.cbr.ru/analytics/analytics\_nfo/*
16. Обзор ключевых показателей страховщиков по итогам 2018 года // Центральный банк Российской Федерации. *https://www.cbr.ru/analytics/analytics\_nfo/*
17. Исаева Ш.М. Условия обеспечения финансовой устойчивости страховых организаций // Современные проблемы науки и образования. – 2015. - № 2-3. - С. 124.
18. Пайевска-квасьны Р. Кречманьска-гиголь К. Европейский и польский страховые рынки в условиях финансового кризиса // Эффективное антикризисное управление. – 2012. - №5. - С. 84-91.
19. Указание Банка России "О предельных размерах базовых ставок страховых тарифов и коэффициентах страховых тарифов, требованиях к структуре страховых тарифов, а также порядке их применения страховщиками при определении страховой премии по обязательному страхованию гражданской ответственности владельцев транспортных средств" от 19 сентября 2014 г. № 3384-У
20. Минаков В.Ф., Радченко М.В., Сингелейцев М.В. Обеспечение финансовой устойчивости и платежеспособности страховых организаций // Современные проблемы науки и образования. – 2014. – № 6.;
21. Закон РФ от 27.11.1992 N 4015-1 (ред. от 28.11.2018) "Об организации страхового дела в Российской Федерации".
22. Базы данных СПАРК // *http://www.spark-interfax.ru/*
23. Архивы ЦБ // *http://lks.fcsm.ru/publication*
24. Webb, A., 2002. Statistical pattern recognition (2nd edn.). New York: Wiley.
25. Требования к финансовой устойчивости страховой компании, Банк ВТБ 24 (ПАО)
26. Требования к финансовой устойчивости страховых компаний, осуществляющих страхование иное, чем страхование жизни. Приложение 1 к Положению о порядке проведения проверки страховых организаций, осуществляющих страхование рисков Клиентов ООО «АТБ» Банк по кредитным сделкам.
27. А.Д. Шеремет, Е.В. Негашев. Методика финансового анализа деятельности коммерческих организаций // - 2-e изд., перераб. и доп. - М.: НИЦ Инфра-М, 2013. - 208 с.
28. Копбаева Л.Т Финансовая устойчивость предприятия: оценка и принятие решений Экономические науки. Современное состояние и перспективы развития: сборник статей Международной научной студенческой конференции. –2016. – 35-41 с.
29. Данилова Н. Л. Сущность и проблемы анализа финансовой устойчивости коммерческого предприятия // Концепт. – 2014.
30. И.В. Пискунова Оценка финансовой устойчивости предприятия в современных условиях хозяйствования // Вопросы экономики и права. -2011. - № 3. - С. 304-310.
31. Буркальцева Д.Д., Блажевич О.Г. и Чередниченко М.С Оценка финансовой устойчивости бизнеса: теоретические аспекты // Science time. -2016 - С. 96-102.
32. Куликов С.В. Методика финансового анализа страховой организации // НГУЭУ, 2005.
33. Yan C., Wang L., Liu W., Qi M., 2017, Financial early warning of non–life insurance company based on RBF neural network optimized by genetic algorithm, . Concurrency Computat: Pract Exper. 2017:e4343.
34. Sabry F., Hrycay W.P., 2014, An Economist's View of Market Evidence in Valuation and Bankruptcy Litigation, NERA Consulting.
35. Rauch J., Wende S., 2015, Solvency prediction for property-liability insurance companies: Evidence from the financial crisis, The Geneva Papers on Risk and Insurance-Issues, 40: 47–65.
36. Объединенные и сводные данные отчетности субъектов страхового дела // <https://www.cbr.ru/finmarket/account/account_repor_insure/archive/>
37. Выдержки из информационного письма Службы Банка России по финансовым рынкам от 25.12.2013 № 50-13-СШ-12/14987 «Об условиях и форматах представления страховщиками и страховыми брокерами отчетности в виде электронных документов».

# **9. Приложение**

**Код Программы на языке программирования Python.**

# загрузка библиотек

import xml.etree.ElementTree as ET

import glob,os

import pandas as pd

from pandas import to\_datetime

import numpy as np

import matplotlib.pyplot as plt

import datetime as dt

import math

from sklearn import preprocessing

from sklearn import metrics

from sklearn.metrics import roc\_auc\_score

from sklearn.metrics import roc\_curve, auc

from sklearn.linear\_model import LogisticRegression

import statsmodels.api as sm

import scipy.stats as stats

from statsmodels.formula.api import ols

from sklearn.linear\_model import LogisticRegressionCV

from sklearn.feature\_selection import RFECV

from sklearn.model\_selection import StratifiedKFold

from sklearn.ensemble import RandomForestClassifier

from statsmodels.stats.outliers\_influence import variance\_inflation\_factor

from statsmodels.tools.tools import add\_constant

import itertools

from sklearn.tree import DecisionTreeClassifier

# путь к папке

ws = "D:\\Documents\\strahov\\"

########################

# Парсинг xml файлов финансовой отчетности 2012-2016 годов

# (Сбор и конвертация финансовых отчетов из xml в сводную таблицу.)

########################

#название папок с архивами

folders = ['xml\_2012\\', 'xml\_2013\\', 'xml\_2014\\','xml\_2015\\', 'xml\_2016\\']

# Формы отчетности

forms = ['1','2']

dfcols = ['id', 'period', 'form','line', 'col', 'value']

df = pd.DataFrame(columns=dfcols)

### Извлечение данных из отчетности за 2011 - 2016 года

for k in folders:

for i in os.listdir(path=ws+k):

for j in forms:

tree = ET.parse(ws+k+i+"\\"+j+".xml")

root = tree.getroot()

period=root.attrib.get('period')

comp\_id=root.attrib.get('id')

for child in root:

for child1 in child:

for child2 in child1:

for child3 in child2:

df = df.append(pd.Series([comp\_id, period, j, child2.attrib.get('id'), child3.attrib.get('id'), child3.text ], index=dfcols), ignore\_index=True)

df.to\_csv(ws+'df\_all.csv', index = False, sep = ';')

df = pd.read\_csv(ws+'df\_all.csv', sep = ';')

# извлечение информации по 2011 годe за счет 5-ой колонки в отчетности следующего года

df['col']=pd.to\_numeric(df['col'])

df = df[(df['col'] == 4)|((df['col'] == 5)&(df['period'] =='31.12.2012'))]

df['period'] = np.where((df['col'] == 5)&(df['period'] =='31.12.2012'),'31.12.2011',df['period'])

df['id']=pd.to\_numeric(df['id'])

df['value'] = df['value'].map(lambda x: ''.join(str(x).split()))

df['value'] = np.where(df['value']!="nan", df['value'], np.nan)

df['value']=pd.to\_numeric(df['value'].str.replace(" ",""))

# трансформация таблицы со значениями по всем финансовым факторам в одной колонке в таблицу с множеством колонок различных финансовых переменных

df0=df[['id','period']].drop\_duplicates()

for j in df['form'].drop\_duplicates():

line\_num=df[df['form']==j]['line'].drop\_duplicates()

for i in line\_num:

df2=df[(df['line']==i)&(df['form']==j)][['id', 'period' ,'value']].drop\_duplicates()

df0 = df0.merge(df2.rename(index=str, columns={'value' : str(j)+"\_"+str(i), }),how='left', on = ['id','period'])

del df2

# Извлечение года из даты

df0['year'] = pd.to\_numeric(df0['period'].str.replace('31.12.',''))

# сохранение итога

df0.to\_csv(ws+'df\_all\_2.csv', index = False, sep = ';')

########################

# формирование финансовых относительных факторов 2012-2016 годов

########################

for i in list(df):

df[i] = df[i].fillna(0)

df['1\_1300'] = np.where(df['1\_1300']<=0,df['1\_1000'],df['1\_1300'])

df['1\_1000']= np.where(df['1\_1000']<=0,df['1\_1300'],df['1\_1000'])

df['1\_2200'] = np.where(df['1\_2200']<=0, df['1\_2000'] - df['1\_2100'],df['1\_2200'])

df['1\_2100'] = np.where(df['1\_2100']<=0, df['1\_2000'] - df['1\_2200'],df['1\_2100'])

col = ['2\_2120', '2\_2200', '2\_2210', '2\_2220', '2\_2500', '2\_2600', '2\_2610', '2\_2620',

'2\_2800', '2\_2920', '2\_3100', '2\_3300', '2\_3500', '2\_1400', '2\_1410', '2\_1600', '2\_1800', '2\_1300','2\_1120']

for i in col:

df[i] = -df[i].abs()

#### Generate parametres for model

df['fs1'] = (df['1\_2100'] )/(df['1\_2200']+df['1\_2100'])

df['fs2'] = df['1\_2100']/(df['1\_2210'] + df['1\_2220'] - df['1\_1230'] - df['1\_1240'])

df['fs3'] = (df['1\_2200'] - df['1\_2210'] - df['1\_2220'] - df['1\_2280']) / df['1\_2000']

df['fs4'] = df['1\_2200']/df['1\_2100']

df['fs5'] = df['1\_2220']/df['1\_1300']

df['fs6'] = (df['2\_2100'] + df['2\_1100']) / (df['1\_2220'] + df['1\_2210'])

#df['fs7'] = df['1\_1300'].map(lambda x: math.log(x))

#df['fs8'] = np.where((df['2\_2100'] + df['2\_1100'])>0,(df['2\_2100'] + df['2\_1100']),np.nan)

#df['fs8'] = df['fs8'].map(lambda x: math.log(x))

df['pr1'] = (df['2\_3400'] ) / (df['2\_1100'] + df['2\_1200'] + df['2\_1300'] +

df['2\_1700'] + df['2\_1800'] + df['2\_2110'] + df['2\_2700'] + df['2\_2800'] + df['2\_2910'] +

df['2\_2920'] + df['2\_3200'] + df['2\_3300'])

df['pr2'] = df['2\_3400'] / df['1\_2100']

df['pr3'] = df['2\_3000'] / df['1\_2100']

df['pr4'] = df['2\_3000'] / (df['2\_1110'] + df['2\_2110'])

df['pr5'] = df['2\_3000'] / df['1\_1300']

df['pr6'] = df['2\_3400'] / df['1\_1300']

df['unpr1'] = -(df['2\_2200'] + df['2\_1400']) /(df['2\_2100'] + df['2\_1100'])

df['unpr2'] = -(df['2\_1600'] + df['2\_1800'] + df['2\_1700'] + df['2\_2220'] + df['2\_2600'] +

df['2\_2920'] + df['2\_2910'] + df['2\_3100'] + df['2\_3300'] + df['2\_3200']) / (df['2\_1100'] + df['2\_2100'] )

df['unpr3'] = -(df['2\_1400'] + df['2\_1500'] + df['2\_1600'] + df['2\_2200'] + df['2\_2600'] +

df['2\_3100']) / (df['2\_2100'] + df['2\_1100'])

df['re1'] = df['1\_1240'] / df['1\_2220']

#df['re2'] = (-(df['2\_2120'] + df['2\_1120']) / (df['2\_1100'] + df['2\_2100']))

df['re2'] = (-(df['2\_2120'] + df['2\_1120']) / (df['2\_1110'] + df['2\_2110']))

#df['re3'] = - (df['2\_1420'] + df['2\_2230'])/(df['2\_1400'] + df['2\_2200'])

df['re3'] = - (df['2\_1420'] + df['2\_2230'])/(df['2\_1410'] + df['2\_2210'])

df['inv'] = (df['1\_1140'] + df['1\_1270'] ) / ( df['1\_2220'] - df['1\_1240'])

df['liq1'] = (df['1\_1140'] + df['1\_1270']) / (df['1\_2200'] - df['1\_1230'] - df['1\_1240'] - df['1\_2280'])

df['liq2'] = (df['1\_1140'] + df['1\_1270']) / df['1\_1300']

df['solv'] = -(df['2\_1100'] + df['2\_2100']) / (df['2\_1400'] + df['2\_1600'] + df['2\_1800'] + df['2\_2200'] + df['2\_2600'] + df['2\_2920'] + df['2\_3100'])

param = ['fs1', 'fs2', 'fs3', 'fs4', 'fs5', 'fs6', #'fs7', 'fs8',

'pr1', 'pr2', 'pr3', 'pr4', 'pr5', 'pr6', 'unpr1',

'unpr2', 'unpr3', 're1', 're2','re3', 'inv', 'liq1', 'liq2', 'solv']

df['prem'] = df['2\_2100'] + df['2\_1100']

df['pay'] = df['2\_2210'] + df['2\_1410']

df['act'] = df['1\_1300']

########################

# парсинг xml-файлов годовой финансовой отчетности 2017 года с новым форматом отчетности

########################

folders = ['xml\_2017\_msfo\\']

forms = ['39\_125','39\_126']

dfcols = ['id', 'period', 'form','line', 'col', 'value']

df = pd.DataFrame(columns=dfcols)

for k in folders:

for i in os.listdir(path=ws+k):

for j in forms:

tree = ET.parse(ws+k+i+"\\"+j+".xml")

root = tree.getroot()

period=root.attrib.get('period')

comp\_id=root.attrib.get('id')

for child in root:

for child1 in child:

if child1.attrib.get('id')!='1c':

for child2 in child1:

df = df.append(pd.Series([comp\_id, period, child.attrib.get('id'), child2.attrib.get('id'), child2[0].attrib.get('id'), child2[0].text ], index=dfcols), ignore\_index=True)

df['col']=pd.to\_numeric(df['col'])

df = df[df['col']== 4]

df['id']=pd.to\_numeric(df['id'])

df['value'] = df['value'].map(lambda x: ''.join(str(x).split()))

df['value'] = np.where(df['value']!="nan", df['value'], np.nan)

df['value']=pd.to\_numeric(df['value'].str.replace(" ",""))

df\_39 = df[df['form'].isin(['39\_125','39\_126'])].drop\_duplicates()

df\_9 = df[df['form'].isin(['9\_125','9\_126'])].drop\_duplicates()

df\_39['form'] = np.where(df\_39['form'] == '39\_125', '1', '2')

forms = ['1','2']

df1 = df\_39[['id', 'period', 'form', 'line', 'value']].drop\_duplicates()

df0=df1[['id','period']].drop\_duplicates()

for j in forms:

line\_num=df1[df1['form']==j]['line'].drop\_duplicates()

for i in line\_num:

df2=df1[(df1['line']==i)&(df1['form']==j)][['id', 'period' ,'value']].drop\_duplicates()

df0 = df0.merge(df2.rename(index=str, columns={'value' : j+"\_"+str(i), }),how='left', on = ['id','period'])

df0['year'] = pd.to\_numeric(df0['period'].str.replace('31.12.',''))

df0['life'] = np.where(~df0['2\_1'].isnull(), 1, 0)

for i in list(df0):

df0[i] = df0[i].fillna(0)

########################

# Формирование относительных финансовых факторов по 2017 году

########################

df0['fs1'] = df0['1\_51'] / (df0['1\_40'] + df0['1\_51'])

df0['fs2'] = df0['1\_51'] / (df0['1\_30'] + df0['1\_33'] - df0['1\_9'] - df0['1\_11'])

df0['fs3'] = (df0['1\_40'] - df0['1\_30'] - df0['1\_33'] ) / df0['1\_52']

df0['fs4'] = df0['1\_40'] / df0['1\_51']

df0['fs5'] = df0['1\_33'] / df0['1\_23']

df0['fs6'] = (df0['2\_1'] + df0['2\_8']) / (df0['1\_33'] + df0['1\_30'])

df0['prem'] = df0['2\_1'] + df0['2\_8']

df0['pay'] = df0['2\_9.1'] + df0['2\_2.1']

df0['pr1'] = (df0['2\_30']) / (df0['2\_1.1'] + df0['2\_5'] + df0['2\_6'] + df0['2\_8.1'] + df0['2\_22']

+ df0['2\_12'] + df0['2\_13'] + df0['2\_27'] + df0['2\_28'])

df0['pr2'] = df0['2\_30'] / df0['1\_51']

df0['pr3'] = df0['2\_54'] / df0['1\_51']

df0['pr4'] = df0['2\_54'] / (df0['2\_1.1'] + df0['2\_8.1'])

df0['pr5'] = df0['2\_54'] / df0['1\_23']

df0['pr6'] = df0['2\_30'] / df0['1\_23']

df0['unpr1'] = (df0['2\_9'] + df0['2\_2']) / (df0['2\_8'] + df0['2\_1'])

df0['unpr2'] = - (df0['2\_4'] + df0['2\_5'] + df0['2\_6'] + df0['2\_9.2'] + df0['2\_10'] + df0['2\_12']

+ df0['2\_13'] + df0['2\_23'] + df0['2\_27'] + df0['2\_28']) / (df0['2\_8'] + df0['2\_1'])

df0['unpr3'] = - (df0['2\_2'] + df0['2\_3'] + df0['2\_4'] + df0['2\_9'] + df0['2\_10'] +

df0['2\_23']) / (df0['2\_8'] + df0['2\_1'])

df0['re1'] = df0['1\_11'] / df0['1\_33']

df0['re2'] = (df0['2\_8.2'] + df0['2\_1.2']) / (df0['2\_1.1'] + df0['2\_8.1'])

df0['re3'] = (df0['2\_2.2'] + df0['2\_9.3']) / (df0['2\_2'] + df0['2\_9.1'])

df0['inv'] = (df0['1\_1'] + df0['1\_2'] + df0['1\_3'] + df0['1\_4'] + df0['1\_5']) / (df0['1\_33'] - df0['1\_11'])

df0['liq1'] = (df0['1\_1'] + df0['1\_2'] + df0['1\_3'] + df0['1\_4'] + df0['1\_5']) / (df0['1\_40'] - df0['1\_9'] - df0['1\_11'] )

df0['liq2'] = (df0['1\_1'] + df0['1\_2'] + df0['1\_3'] + df0['1\_4'] + df0['1\_5']) / df0['1\_23']

df0['solv'] = (df0['2\_8'] + df0['2\_1']) / (df0['2\_2'] + df0['2\_4'] + df0['2\_6'] + df0['2\_9'] + df0['2\_10'] + df0['2\_13'] + df0['2\_23'])

df0['act'] = df0['1\_23']

df0.to\_csv(ws+'df\_2017.csv', sep = ';')

df0 = pd.read\_csv(ws+'df\_2017.csv', sep = ';')

df17 = df0[['id','year','prem','pay','act'] + param]

df17.to\_csv(ws+'df\_2017\_short.csv', sep = ';')

########################

# Формирование единой база 2012-2017 года и добавление переменных, #отвечающих за динамику и дамми-переменных

########################

# объединение данных за 2011-2016 и 2017 год

df = pd.concat([df[['id','year','prem','pay','act'] + param],df17[['id','year','prem','pay','act'] + param]],ignore\_index=True)

# вычисление факторов изменения премий и выплат

df\_lag = df[['id','year','prem','pay']].copy()

for i in ['prem','pay']:

df\_lag = df\_lag.rename(columns={i:i +'\_lag'})

df\_lag['year'] = df['year'] + 1

df = df.merge(df\_lag, how = 'left', on = ['id','year'])

for i in ['prem','pay']:

df['delta\_' + i] = (df[i] - df[i + '\_lag'])/df[i]

del df\_lag

# удаление наблююдений по 2011 и 2010 годам

df1 = df[~df['year'].isin([2010,2011])].drop\_duplicates()

# загрузка базы дефолтов

default = pd.read\_excel(ws + 'dead\_alive.xlsx')

default['def'] = np.where(default['default\_date'].isnull(),0,1)

default['year'] = default['default\_date'].map(lambda x: x.year) - 1

default = default[['id','year','def']].drop\_duplicates()

# проставление маркеров дефолта в нашей выборке

df1 = df1.merge(default[['id','year','def']].rename(columns = {'def':'def\_1'}),how='left',on=['id','year'])

df1['def\_1'] = np.where(df1['def\_1'].isnull(),0,df1['def\_1'])

# сохранение полученной выборки

df1.to\_csv(ws+'data\_all\_17\_new.csv', index = False)

df1 = pd.read\_csv(ws+'data\_all\_17\_new.csv')

df1['pay'] = -df1['pay']

param = ['fs1','fs2','fs3','fs4','fs5','fs6','pr1','pr2', 'pr3', 'pr4', 'pr5','pr6',

'unpr1', 'unpr2', 'unpr3', 're1', 're2', 're3', 'inv', 'liq1', 'liq2', 'solv',

'prem', 'pay', 'delta\_prem', 'delta\_pay', 'act',

'insure\_sum\_new', 'num\_zayav', 'num\_ureg', 'num\_ref', 'num\_all\_new', 'num\_all',

'num\_ref\_div\_zayav', 'num\_ureg\_div\_zayav', 'num\_ureg\_ref\_div\_zayav', 'num\_zayav\_div\_all',

'prem\_div\_insure\_sum\_new', 'num\_all\_div\_new']

p\_add1 = ['delta\_prem','div\_num\_all','div\_insure\_sum\_new','num\_ref',

'div\_num\_zayav','div\_act','div\_pay','num\_ref\_div\_zayav',

'prem\_div\_insure\_sum\_new']

param\_new = ['fs1', 'fs2', 'fs4', 'liq1','liq2','pr1','pr2', 'pr4', 'pr6',

're2', 'solv', 'unpr1', 'unpr2', 'unpr3', 'life']

df1 = df1[['id','year','def\_1']+param]

filling\_missings = True

# добавление переменной дамми-года

df1 = pd.concat([df1, pd.get\_dummies(df1['year'])], axis=1)

param\_year = list(pd.get\_dummies(df1['year']))

# добавление дамми-переменных видов страховой деятельности

param\_kind = [ 'life', 'medical', 'medical\_voluntary', 'avto',

'exceptlife\_other', 'reinsurance\_in', 'reinsurance\_out',

'liability', 'mutual']

df = pd.read\_csv(ws+'kinds\_full.csv')

df1 = df1.merge(df[['id','year']+param\_kind],how='left',on = ['id','year'])

d = pd.read\_csv(ws+'kind.csv')[['id','year','ca']]

df1 = df1.merge(d,how='left',on=['id','year'])

param = param+['ca']

del d,df

# добавление готового региона регистрации из баз спарка

ws1 = ws + 'spark/region/'

sp = pd.read\_excel(ws1 + os.listdir(path=ws1)[0], skiprows=[0,1,2])

sp['year'] = int(os.listdir(path=ws1)[0].replace('.xlsx',''))

for i in os.listdir(path=ws1):

if i != os.listdir(path=ws1)[0]:

sp\_new = pd.read\_excel(ws1 + i, skiprows=[0,1,2])

sp\_new['year'] = int(i.replace('.xlsx',''))

sp = pd.concat([sp,sp\_new],ignore\_index=True)

sp = sp.rename(columns={'Год':'year','Регистрационный номер':'id','Регион регистрации':'region'})

del sp\_new

# заполнение оставшихся пропусков по регионам вручную

sp\_add = pd.DataFrame({

'id' : [1,1209,4296,4307,4323,4335,4337,4348,4349,4351],

'region' : ['Москва', 'Москва', 'Вологодская область', 'Республика Татарстан', 'Москва',

'Республика Крым', 'Республика Крым', 'Тюменская область', 'Санкт-Петербург', 'Москва'],

})

# добавление к нашей базе

s = pd.concat([sp[['id','region']],sp\_add],ignore\_index=True)

df1 = df1.merge(s[['id','region']].drop\_duplicates(),how = 'left',on='id')

df1['reg\_mos'] = np.where(df1['region']=='Москва',1,0)

df1.to\_csv(ws + 'data\_normal2.csv',index=False)

########################

# Трансформация факторов

########################

# Заполнение пропусков

if filling\_missings:

for i in list(df1):

df1[i] = np.where(df1[i] == float('-inf'), df1[df1[i] != float('-inf')][i].min(),

np.where(df1[i] == float('inf'), df1[df1[i] != float('inf')][i].max(),

np.where(~df1[i].isnull(), df1[i], 0)))

# df1[i].fillna(0)#df1[i].median())

else:

#param = param\_new + p\_add1

df1 = df1[['id','year','def\_1']+param]

for i in list(df1):

df1 = df1[~df1[i].isnull()]

# Нормализация финансовых переменных

for i in param:

if i!='life':

median = df1[i].median()

df1[i] = df1[i].fillna(median)

perc = np.percentile(df1[i], 95)

slope=-(math.log(1/0.95-1)/(perc-median))

df1[i] = df1[i].map(lambda x: 1/(1+math.exp(-slope\*(x-median)))

if -slope\*(x-median)<700 else 0 )

# mean = df1[i].mean()

# std = df1[i].std()

# df1[i]=(df1[i]-mean)/std

del median,perc,slope

# Итоговые факторы

param\_year = ['2012','2013','2014','2015','2016','2017']

par = param + param\_kind + param\_year

##############

# Однофакторный анализ

##############

# Gini по отдельным переменным

def Gini\_fun(df1,par):

dfcols\_g = ['param','sg', 'gini']

Gini = pd.DataFrame(columns=dfcols\_g)

for i in par:

pred = df1[[i]]

roc\_auc = roc\_auc\_score(df1[['def\_1']], pred)

gini = abs(2\*roc\_auc - 1)

sg = np.sign(2\*roc\_auc - 1)

Gini = Gini.append(pd.Series([i,sg, gini], index=dfcols\_g), ignore\_index=True)

del roc\_auc,gini,sg,dfcols\_g,pred

return Gini

Gini = Gini\_fun(df1,par)

# T-test по отдельным переменным

dfcols\_t = ['param', 'P-value']

Test = pd.DataFrame(columns=dfcols\_t)

for k in par:

X = df1[[k] + ['def\_1']].copy()

st, pval = stats.f\_oneway(X[X['def\_1']==1][k], X[X['def\_1']==0][k])

Test = Test.append(pd.Series([k, pval], index=dfcols\_t), ignore\_index=True)

T = Gini.merge(Test,how = 'left',on='param')

writer = pd.ExcelWriter(ws + 'Gini\_Test\_add1.xlsx')

T.to\_excel(writer, index = False)

writer.save()

# переменные, у которых влияние правильное

par\_new = ['fs3', 'fs6', 'pr1', 'pr2', 'pr3', 'pr4', 'pr5', 'pr6', 'unpr1',

'unpr2', 'unpr3', 're1', 're2', 're3', 'inv', 'liq1', 'liq2',

'solv', 'prem', 'delta\_prem', 'act', 'num\_all\_new', 'num\_all',

'prem\_div\_insure\_sum\_new', 'ca']

par\_not\_u = ['fs3', 'pr2', 'pr3', 'pr5', 'pr6', 'unpr1', 'unpr2',

'unpr3', 're3', 'solv', 'prem', 'act', 'num\_all\_new',

'num\_all', 'prem\_div\_insure\_sum\_new', 'ca']

par\_u = ['fs6', 'pr1', 'pr4', 're1', 're2', 'inv', 'liq1', 'liq2', 'delta\_prem']

# добавление квадратов переменных

dfcols\_g = ['param', 'gini','gini\_u']

Gini\_u = pd.DataFrame(columns=dfcols\_g)

p2 =[]

for i in par\_u:

p2 = p2 + [i +'^1\_2']

X\_train[i+'^2'] = X\_train[i]\*\*2

logit\_model=sm.Logit(y\_train,X\_train[[i,i+'^2']])

result=logit\_model.fit()

#print(result.summary2())

df1[i+'^2'] = df1[i]\*df1[i]

df1[i +'^1\_2'] = result.predict(df1[[i,i+'^2']])

roc\_auc = roc\_auc\_score(df1[['def\_1']], df1[i +'^1\_2'])

gini\_u = abs(2\*roc\_auc - 1)

Gini\_u = Gini\_u.append(pd.Series([i,Gini[Gini['param']==i]['gini'].mean(), gini\_u], index=dfcols\_g), ignore\_index=True)

del roc\_auc,gini\_u,dfcols\_g

writer = pd.ExcelWriter(ws + 'Gini\_u.xlsx')

Gini\_u.to\_excel(writer, index = False)

writer.save()

# построение графиков roc-кривых

for j in par\_u:

for i in [j,j +'^1\_2']:

y = df1[['def\_1']].copy()

pred = df1[[i]].copy()

fpr, tpr, \_ = metrics.roc\_curve(y, pred, pos\_label=1)

roc\_auc = auc(fpr, tpr)

gini = abs(2\*roc\_auc - 1)

sg = np.sign(2\*roc\_auc - 1)

#Gini = Gini.append(pd.Series([i,sg, gini], index=dfcols\_g), ignore\_index=True)

plt.figure()

lw = 2

plt.plot(fpr, tpr, color='darkorange',

lw=lw, label='ROC curve (area = %0.2f)' % roc\_auc)

plt.plot([0, 1], [0, 1], color='navy', lw=lw, linestyle='--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.xlabel('False Positive Rate')

plt.ylabel('True Positive Rate')

plt.title('Receiver operating characteristic of ' + i)

plt.legend(loc="lower right")

plt.savefig(ws + 'pict/' + i + '.png')

plt.show()

print(gini)

# Матрица корреляций

M = df1[par\_not\_u + p2].corr().abs()

writer = pd.ExcelWriter(ws + 'Corr\_matrix2.xlsx')

M.to\_excel(writer, index = False)

writer.save()

# Удаление сильно коррелируемых переменных

Gini = Gini\_fun(df1,par\_not\_u + p2)

dfcols\_cor = ['param1','param2','Corr', 'gini1','gini2']

M\_cor = pd.DataFrame(columns=dfcols\_cor)

p = (par\_not\_u + p2).copy()

#p = Gini.sort\_values(by = 'gini',ascending=False)['param'].copy()

p1 = []

for i in p:

for j in p:

if (i!=j) and (M[i][j]>0.75):

gini1 = Gini[Gini['param']==i]['gini'].max()

gini2 = Gini[Gini['param']==j]['gini'].max()

if (gini1>=gini2):

print(i,j)

p1 += [j]

M\_cor = M\_cor.append(pd.Series([i,j,M[i][j],gini1,gini2], index=dfcols\_cor), ignore\_index=True)

writer = pd.ExcelWriter(ws + 'Corr\_matrix3.xlsx')

M\_cor.to\_excel(writer, index = False)

writer.save()

p1 = list(set(p1))

p1.remove('act')

for i in p1:

p.remove(i)

df1.to\_csv(ws + 'data\_normal2.csv')

df1 = pd.read\_csv(ws + 'data\_normal2.csv')

########################

# Создание моделей

########################

p = ['fs3','pr2','unpr1','unpr2','re3','solv','act','num\_all','prem\_div\_insure\_sum\_new',

'ca','fs6^1\_2','pr1^1\_2','pr4^1\_2','re1^1\_2','re2^1\_2','liq1^1\_2','liq2^1\_2','delta\_prem^1\_2']

param\_year = ['2012','2013','2014','2015','2016']

param\_kind = [# 'address',

'life', 'medical', 'medical\_voluntary', 'avto',

'exceptlife\_other', 'reinsurance\_in', 'reinsurance\_out',

'liability', 'mutual']

region = ['reg\_mos']

par = p + param\_kind + region + param\_year + ['const'] #+ ['2012-2013','2014-2015']#

#разделение выборки на train и test

df1 = add\_constant(df1)

X\_train = df1[df1['year']!=2017][par].copy()

y\_train = df1[df1['year']!=2017][['def\_1']].copy()

X\_test = df1[df1['year']==2017][par].copy()

y\_test = df1[df1['year']==2017][['def\_1']].copy()

def logit\_fun(p):

logit\_model=sm.Logit(y\_train,X\_train[p])

result=logit\_model.fit()

res = result.summary2()

#print(res)

predicted = result.predict(X\_train[p])

roc\_auc = roc\_auc\_score(y\_true=y\_train, y\_score=predicted)

gini\_tr = abs(2\*roc\_auc - 1)

#print('gini\_train = ', gini\_tr)

predicted = result.predict(X\_test[p])

roc\_auc = roc\_auc\_score(y\_true=y\_test, y\_score=predicted)

gini\_ts = abs(2\*roc\_auc - 1)

#print('gini\_test = ', gini\_ts)

d = pd.DataFrame({'pred':predicted,'fact':y\_test['def\_1']})

d['pred1']=np.where(d['pred']>0.2,1,0)

d['n']=1

d1 = pd.DataFrame({'0\_fact':[d[(d['pred1']==0)&(d['fact']==0)]['n'].sum(),d[(d['pred1']==1)&(d['fact']==0)]['n'].sum()],

'1\_fact':[d[(d['pred1']==0)&(d['fact']==1)]['n'].sum(),d[(d['pred1']==1)&(d['fact']==1)]['n'].sum()]})

print(d1)

return res, gini\_tr, gini\_ts

def logit\_fun\_1(p):

clf = LogisticRegression()#LogisticRegressionCV(random\_state=1)

clf.fit(X\_train[p], y\_train)

predicted = clf.predict\_proba(X\_train[p])

roc\_auc = roc\_auc\_score(y\_true=y\_train, y\_score=predicted)

gini\_tr = abs(2\*roc\_auc - 1)

return gini\_tr, clf.densify(), clf.get\_params()

gini\_tr, densify, params = logit\_fun\_1(par)

def roc\_pic(p):

logit\_model=sm.Logit(y\_train,X\_train[p])

result=logit\_model.fit()

predicted = result.predict(X\_train[p])

fpr\_tr, tpr\_tr, \_ = metrics.roc\_curve(y\_train, predicted, pos\_label=1)

roc\_auc\_tr = auc(fpr\_tr, tpr\_tr)

gini\_tr = abs(2\*roc\_auc\_tr - 1)

predicted = result.predict(X\_test[p])

fpr\_ts, tpr\_ts, \_ = metrics.roc\_curve(y\_test, predicted, pos\_label=1)

roc\_auc\_ts = auc(fpr\_ts, tpr\_ts)

gini\_ts = abs(2\*roc\_auc\_ts - 1)

#Gini = Gini.append(pd.Series([i,sg, gini], index=dfcols\_g), ignore\_index=True)

plt.figure()

lw = 2

plt.plot(fpr\_tr, tpr\_tr, color='darkorange',

lw=lw, label='Gini train = %0.2f' % gini\_tr)

plt.plot(fpr\_ts, tpr\_ts, color='orange',

lw=lw, label='Gini test = %0.2f' % gini\_ts)

plt.plot([0, 1], [0, 1], color='navy', lw=lw, linestyle='--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.xlabel('False Positive Rate')

plt.ylabel('True Positive Rate')

#plt.title('Receiver operating characteristic of ' + i)

plt.legend(loc="lower right")

plt.savefig(ws+ "ROC.png")

plt.show()

#print(gini)

# Логит-модель от всех переменных

res, gini\_tr, gini\_ts = logit\_fun(par)

roc\_pic(par)

#clf = LogisticRegression().fit(X\_train, y\_train)

#clf.predict(X\_test)

# Логит-модель от значимых переменных

par\_new = ['pr2','unpr1','act','fs6^1\_2','pr4^1\_2','re1^1\_2','re2^1\_2','liq2^1\_2','delta\_prem^1\_2',

'life','medical','2012','2013'] +['const']

res, gini\_tr, gini\_ts = logit\_fun(par\_new)

print(res)

print('gini\_train = ', gini\_tr)

print('gini\_test = ', gini\_ts)

roc\_pic(par\_new)

# vif

def vif(p):

X = add\_constant(X\_train[p])

return pd.Series([variance\_inflation\_factor(X.values, i)

for i in range(X.shape[1])],

index=X.columns)

v = vif(par\_new)

writer = pd.ExcelWriter(ws + 'vif.xlsx')

v.to\_excel(writer)

writer.save()

# RFECV

def RFE\_feature\_selection(clf\_lr,n):

rfecv = RFECV(estimator=clf\_lr, step=1,verbose=0, cv=StratifiedKFold(n), scoring='roc\_auc')

rfecv.fit(X\_train, y\_train)

print("Optimal number of features : %d" % rfecv.n\_features\_)

# Plot number of features VS. cross-validation scores

f, ax = plt.subplots(figsize=(14, 9))

plt.xlabel("Number of features selected")

plt.ylabel("Cross validation score (roc\_auc)")

plt.plot(range(1, len(rfecv.grid\_scores\_) + 1), rfecv.grid\_scores\_)

plt.savefig(ws+ "fselect.png")

plt.show()

mask = rfecv.get\_support()

X = X\_train.ix[:, mask]

par\_new = list(X)

print(rfecv.grid\_scores\_)

res, gini\_tr, gini\_ts = logit\_fun(par\_new)

print(res)

print('gini\_train = ', gini\_tr)

print('gini\_test = ', gini\_ts)

return list(X)

clf\_lr = LogisticRegression()#LogisticRegressionCV(random\_state=1)

clf\_lr.fit(X\_train, y\_train)

RFE\_feature\_selection(clf\_lr,5)

RFE\_feature\_selection(clf\_lr,2)

# Перебор различных подмножеств факторов

# 8 и 9

fl = list(itertools.combinations(p,8))+list(itertools.combinations(p,9))

flk = list(itertools.combinations(param\_kind,1))+list(itertools.combinations(param\_kind,2))

fly = [['2012','2013', 'const']]#fly1 + fly2

dfcols\_g = ['param','num', 'gini']

M = pd.DataFrame(columns=dfcols\_g)

n=0

for i1 in fl:

# for i2 in flk:

# for i3 in fly:

i = list(i1)#+i2+i3

n=n+1

X = X\_train[i].copy()

y = y\_train[['def\_1']].copy()

logit\_model=sm.Logit(y,X[i])

result=logit\_model.fit()

#print(result.summary2())

predicted = result.predict(X[i])

roc\_auc = roc\_auc\_score(y\_true=y\_train, y\_score=predicted)

gini = abs(2\*roc\_auc - 1)

print('gini = ', gini)

print('num = ', n)

p1 = ';'

for j in list(np.arange(len(i))):

p1 = p1 + i[j] + ";"

M = M.append(pd.Series([p1, len(i),gini], index=dfcols\_g), ignore\_index=True)

M.to\_csv(ws + 'M\_8\_or\_9\_param.csv',index=False)

(M['gini']>0.425).sum()

M[M['gini']>0.35]['param']

dfcols\_g1 = ['param','num', 'gini1','gini2']

M1 = pd.DataFrame(columns=dfcols\_g1)

n=0

for l in M[M['gini']>0.425]['param']:

i1 = l.split(';')

i1.remove('')

i1.remove('')

gini1 = M[(M['param']==l)]['gini'].mean()

for i2 in flk:

for i3 in fly:

i = i1+i2+i3

n=n+1

X = X\_train[i].copy()

y = y\_train[['def\_1']].copy()

logit\_model=sm.Logit(y,X[i])

result=logit\_model.fit()

#print(result.summary2())

predicted = result.predict(X[i])

roc\_auc = roc\_auc\_score(y\_true=y\_train, y\_score=predicted)

gini = abs(2\*roc\_auc - 1)

print('gini = ', gini)

print('num = ', n)

p = ';'

for j in list(np.arange(len(i))):

p1 = p1 + str(i[j]) + ";"

M1 = M1.append(pd.Series([p1, len(i), gini1, gini], index=dfcols\_g1), ignore\_index=True)

M1.to\_csv(ws + 'M\_8\_or\_9\_param\_with\_kind\_year.csv',index=False)

dfcols\_g2 = ['param','num', 'gini\_tr','gini\_ts']

M2 = pd.DataFrame(columns=dfcols\_g2)

(M1['gini2']>0.55).sum()

for l in M1[M1['gini2']>0.55]['param']:

l1 = l.split(';')

l1.remove('')

l1.remove('')

X = X\_train[l1].copy()

y = y\_train[['def\_1']].copy()

logit\_model=sm.Logit(y,X[l1])

result=logit\_model.fit()

predicted = result.predict(X\_test[l1])

roc\_auc = roc\_auc\_score(y\_true=y\_test, y\_score=predicted)

gini = abs(2\*roc\_auc - 1)

print('gini = ', gini)

M2 = M2.append(pd.Series([l, len(l1), M1[M1['param']==l]['gini2'].mean(), gini], index=dfcols\_g2), ignore\_index=True)

l = ';pr2;unpr1;num\_all;prem\_div\_insure\_sum\_new;fs6^1\_2;pr1^1\_2;pr4^1\_2;liq2^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const;'

l1 = l.split(';')

l1.remove('')

l1.remove('')

l1.remove('prem\_div\_insure\_sum\_new')

l1.remove('pr1^1\_2')

res, gini\_tr, gini\_ts = logit\_fun(l1)

print(res)

print('gini\_train = ', gini\_tr)

print('gini\_test = ', gini\_ts)

#### 5,6,7

fl = list(itertools.combinations(p,5))+list(itertools.combinations(p,6))+list(itertools.combinations(p,7))

flk = ['life','medical']

fly = ['2012','2013', 'const']#fly1 + fly2

def pereb(fl,flk,fly):

dfcols\_g = ['param','num', 'gini\_tr','gini\_ts']

M3 = pd.DataFrame(columns=dfcols\_g)

n=0

for i1 in fl:

# for i2 in flk:

# for i3 in fly:

i = list(i1) + flk + fly#+i2+i3

n=n+1

res, gini\_tr, gini\_ts = logit\_fun(i)

print('gini\_ts = ', gini\_ts)

print('num = ', n)

p = ';'

for j in list(np.arange(len(i))):

p = p + i[j] + ";"

M3 = M3.append(pd.Series([p, len(i),gini\_tr, gini\_ts], index=dfcols\_g), ignore\_index=True)

return M3

M3.to\_csv(ws + 'M\_5\_or\_6\_or\_7\_param\_with\_kind\_year.csv',index=False)

writer = pd.ExcelWriter(ws + 'M\_5\_or\_6\_or\_7\_param\_with\_kind\_year.xlsx')

M3.to\_excel(writer)

writer.save()

for l in M4[(M4['gini\_tr']>=0.555)&(M4['gini\_ts']>0.60)&(M4['num']==13)]['param']:

#l = ';pr2;unpr1;num\_all;fs6^1\_2;pr4^1\_2;re1^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const;'

l1 = l.split(';')

l1.remove('')

l1.remove('')

res, gini\_tr, gini\_ts = logit\_fun(l1)

print(res)

print('gini\_train = ', gini\_tr)

print('gini\_test = ', gini\_ts)

roc\_pic(l1)

fl = list(itertools.combinations(p,8))+list(itertools.combinations(p,9))

M4 = pereb(fl,flk,fly)

writer = pd.ExcelWriter(ws + 'M\_8\_or\_9\_param\_with\_kind\_year.xlsx')

M3.to\_excel(writer)

writer.save()

M4.to\_csv(ws + 'M\_8\_or\_9\_param\_with\_kind\_year.csv',index=False)

M3[(M3['gini\_tr']>0.55)].max().shape

M31 = M3[(M3['gini\_tr']>0.55)&(M3['gini\_ts']>0.56)].shape

writer = pd.ExcelWriter(ws + 'M\_8\_or\_9\_param\_with\_kind\_year\_short.xlsx')

M41.to\_excel(writer)

writer.save()

M41 = M4[(M4['gini\_tr']>=0.555)&(M4['gini\_ts']>0.61)&(M4['num']==13)].shape

# Машинное обучение

DecisionTree = DecisionTreeClassifier(criterion='gini', max\_depth = 20, min\_samples\_leaf = 1,min\_samples\_split=4 , random\_state=0)

result=DecisionTree.fit(X\_train[par], y\_train)

print(DecisionTree.feature\_importances\_)

predicted = result.predict(X\_train[par])

roc\_auc = roc\_auc\_score(y\_true=y\_train, y\_score=predicted)

print('gini\_tr = ', abs(2\*roc\_auc - 1))

predicted = result.predict(X\_test[par])

roc\_auc = roc\_auc\_score(y\_true=y\_test, y\_score=predicted)

print('gini\_test = ', abs(2\*roc\_auc - 1))

# вывод самого дерева в напечатанном виде

dt\_plot(DecisionTree,X\_train, X\_test, y\_train, y\_test)

# Логит-модель с предсказаниями только в виде "0" и "1"

def logit\_fun2(p):

logit\_model=sm.Logit(y\_train,X\_train[p])

result=logit\_model.fit()

res = result.summary2()

#print(res)

predicted\_tr = result.predict(X\_train[p])

roc\_auc = roc\_auc\_score(y\_true=y\_train, y\_score=predicted\_tr)

gini\_tr = abs(2\*roc\_auc - 1)

print('gini\_train = ', gini\_tr)

predicted\_ts = result.predict(X\_test[p])

roc\_auc = roc\_auc\_score(y\_true=y\_test, y\_score=predicted\_ts)

gini\_ts = abs(2\*roc\_auc - 1)

print('gini\_test = ', gini\_ts)

predicted\_tr = add\_constant(predicted\_tr)

predicted\_ts = add\_constant(predicted\_ts)

dfcols\_a = ['a', 'gini']

A = pd.DataFrame(columns=dfcols\_a)

# Поиск наилучшей границы P\*

for a in list(predicted\_tr[0].drop\_duplicates()) + [0.99]:

predicted\_tr[1] = np.where( predicted\_tr[0]>=a,1,0) #result.predict(predicted\_tr)

roc\_auc = roc\_auc\_score(y\_true=y\_train, y\_score=predicted\_tr[1])

gini\_tr\_new = abs(2\*roc\_auc - 1)

#print('gini\_train\_new = ', gini\_tr\_new)

A = A.append(pd.Series([a,gini\_tr\_new], index=dfcols\_a), ignore\_index=True)

a = A[A['gini']==A['gini'].max()]['a'].mean()

print("a = ", a)

# Построение матриц кол-ва попаданий и ошибок первого и второго рода

d = pd.DataFrame({'pred':predicted\_tr[0],'fact':y\_train['def\_1']})

d['pred1']=np.where(d['pred']>a,1,0)

roc\_auc = roc\_auc\_score(y\_true=y\_train, y\_score=d['pred1'])

gini\_ts\_new = abs(2\*roc\_auc - 1)

print('gini\_test\_new = ', gini\_ts\_new)

d['n']=1

d1 = pd.DataFrame({'0\_fact':[d[(d['pred1']==0)&(d['fact']==0)]['n'].sum(),d[(d['pred1']==1)&(d['fact']==0)]['n'].sum()],

'1\_fact':[d[(d['pred1']==0)&(d['fact']==1)]['n'].sum(),d[(d['pred1']==1)&(d['fact']==1)]['n'].sum()]})

print(d1)

d = pd.DataFrame({'pred':predicted\_ts[0],'fact':y\_test['def\_1']})

d['pred1']=np.where(d['pred']>=a,1,0)

roc\_auc = roc\_auc\_score(y\_true=y\_test, y\_score=d['pred1'])

gini\_ts\_new = abs(2\*roc\_auc - 1)

print('gini\_test\_new = ', gini\_ts\_new)

d['n']=1

d['n']=1

d2 = pd.DataFrame({'0\_fact':[d[(d['pred1']==0)&(d['fact']==0)]['n'].sum(),d[(d['pred1']==1)&(d['fact']==0)]['n'].sum()],

'1\_fact':[d[(d['pred1']==0)&(d['fact']==1)]['n'].sum(),d[(d['pred1']==1)&(d['fact']==1)]['n'].sum()]})

print(d2)

return A, a, d1, d2

# Итоговое сравнение моделей

Model = 3.2

if Model==3.2:

l = ';pr2;unpr1;num\_all;fs6^1\_2;pr4^1\_2;re1^1\_2;liq2^1\_2;life;medical;2012;2013;const;'

else

l = ';pr2;unpr1;num\_all;fs6^1\_2;pr4^1\_2;re1^1\_2;delta\_prem^1\_2;life;medical;2012;2013;const;'

l1 = l.split(';')

l1.remove('')

l1.remove('')

A, a, d1, d2 = logit\_fun2(l1)

A = A.sort\_values(by='a')

plt.figure()

lw = 2

plt.plot(A['a'], A['gini'], color='navy',

lw=lw)

plt.xlim([0.0, 1.0])

plt.xlabel('P\*')

plt.ylabel('Gini')

plt.savefig(ws+ "P\_gini.png")

plt.show()

writer = pd.ExcelWriter(ws + 'd1.xlsx')

d1.to\_excel(writer)

writer.save()

# Построение итоговой модели, обученной на всей выборке

logit\_model=sm.Logit(df1['def\_1'],df1[l1])

result=logit\_model.fit()

res = result.summary2()

print(res)

predicted\_tr = result.predict(df1[l1])

roc\_auc = roc\_auc\_score(y\_true=df1[['def\_1']], y\_score=predicted\_tr)

gini = abs(2\*roc\_auc - 1)

print('gini\_train = ', gini)

predicted\_tr = add\_constant(predicted\_tr)

dfcols\_a = ['a', 'gini']

A = pd.DataFrame(columns=dfcols\_a)

# Поиск наилучшей границы P\*

for a in list(predicted\_tr[0].drop\_duplicates()) + [0.99]:

predicted\_tr[1] = np.where( predicted\_tr[0]>=a,1,0) #result.predict(predicted\_tr)

roc\_auc = roc\_auc\_score(y\_true=df1[['def\_1']], y\_score=predicted\_tr[1])

gini\_tr\_new = abs(2\*roc\_auc - 1)

#print('gini\_train\_new = ', gini\_tr\_new)

A = A.append(pd.Series([a,gini\_tr\_new], index=dfcols\_a), ignore\_index=True)

a = A[A['gini']==A['gini'].max()]['a'].mean()

print("a = ", a)

# Построение матриц кол-ва попаданий и ошибок первого и второго рода

d = pd.DataFrame({'pred':predicted\_tr[0],'fact':y\_train['def\_1']})

d['pred1']=np.where(d['pred']>a,1,0)

roc\_auc = roc\_auc\_score(y\_true=df1[['def\_1']], y\_score=d['pred1'])

gini\_ts\_new = abs(2\*roc\_auc - 1)

print('gini\_test\_new = ', gini\_ts\_new)

d['n']=1

d1 = pd.DataFrame({'0\_fact':[d[(d['pred1']==0)&(d['fact']==0)]['n'].sum(),d[(d['pred1']==1)&(d['fact']==0)]['n'].sum()],

'1\_fact':[d[(d['pred1']==0)&(d['fact']==1)]['n'].sum(),d[(d['pred1']==1)&(d['fact']==1)]['n'].sum()]})

print(d1)

# Расчет VIF

X = add\_constant(df1[l1])

vf = pd.Series([variance\_inflation\_factor(X.values, i)

for i in range(X.shape[1])],

index=X.columns)

writer = pd.ExcelWriter(ws + 'vif.xlsx')

vf.to\_excel(writer)

writer.save()

1. В обозначениях формул первое число обозначает номер формы, а вторая – строку в данной форме бухгалтерской отчетности [↑](#footnote-ref-1)