Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра математических методов прогнозирования

Магистерская программа «Логические и комбинаторные методы анализа данных»

Магистерская диссертация

«Методы селективного комбинирования признаковой информации в задаче кредитного скоринга при наличии ограничений»

Работу выполнила
Жосан Юлия Сергеевна

Научный руководитель:
к.ф-м.н., профессор
Красоткина Ольга Вячеславовна

Москва

2017

Оглавление

Аннотация

Задача кредитный скоринга является фундаментальной и одной из самых сложных задач, с которыми приходится сталкиваться финансовым учреждениям. Обычно эта задача заключается в прогнозе вероятности дефолта для клиентов банка. В данной работе предлагается унифицированная процедура, основанная на базе регуляризационной логистической регрессией при условии экспертных ограничений на коэффициенты модели. Экспертные ограничения необходимы при искаженных или неполных входных данных, в результате которых признаки входят в модель с некорректными коэффициентами. Экспериментальные результаты показывают, что предлагаемая структура конкурентоспособна и способна создать более интерпретируемую модель.

1 Введение

В современном мире все больше расширяется и развивается банковская сфера деятельности. Кредитование – самый развитый сегмент в данной области, но при этом он является самой рисковым для банков. Поэтому необходимо организовывать качественное управление кредитными рисками и в частности оценивать надежность или кредитоспособность потенциальных заемщиков. Надежными считаются те заемщики, которые могут в срок полностью рассчитаться с кредитными обязательствами.

Существует большое количество методик решения вышеописанной проблемы. Самой распространённой считается кредитный скоринг. В идеологии скоринговых систем лежит гипотеза, что люди со схожими социальными и поведенческими показателями ведут себя одинаково. Принимая такую гипотезу как факт, мы можем строить различные статистические модели, которые существенно облегчают и повышают эффективность работы аналитиков в рамках любого бизнеса.

Если мы присвоим определенный вес каждой из выбранных демографических характеристик клиента (например: пол, возраст, место проживания, должность, длительность работы в одной организации и пр.), то у нас появляется возможность отнести к группам слабо или сильно соответствующих бизнесу новых и потенциальных клиентов на основе их анкеты. Таким образом, клиенту автоматически, на основе результатов анализа скоринговой системы, присваивается целочисленный ранг, указывающий степень доверия и описывающий уровень внимания, которое следует уделять данному клиенту со стороны бизнеса.

Главной проблемой при решении задачи кредитного скоринга является определение значимости характеризующих заемщика признаков и, соответственно, их весовых коэффициентов в модели. Для определения этих значений модель строится на большой выборке, на которой уже известно, является ли заемщик дефолтным или нет. Обычно каждый заемщик описывается множеством факторов, таких как характеристики из кредитной истории, демографические данные, данные из социальных сетей и от телеком-компаний.

Активное развитие скоринговых систем и их внедрение в деятельность банков России необходима не только для самих банков, как инструмент контроля рисков при выдаче кредита заемщику, но и для заемщиков, так как существенно сокращает время рассмотрения поданной им заявки на кредит. На западном рынке существует много программных решений кредитного скоринга, так как внедрение подобных систем там началось давно. Среди наиболее известных – решения от SAS, SPSS, Experian-Scorex и т.д. Существует так же большое количество российских разработок, но у большинства банком

достаточно сильно отличаются потоки клиентов, поэтому использование готового коробочного решения представляется малоэффективным. В виду этого возникает необходимость самостоятельного построения банком правильной скоринговой модели, которая бы выделяла некредитоспособных заемщиков именно из их потока клиентов.

В рамках представленной работы мы не станем подробно рассматривать иные области применения скоринговых систем, однако стоит также отметить, что они могут быть использованы в маркетинге.

Скоринговые системы могут строиться на множестве различных моделей, например деревья решений, нейронные сети, генетические алгоритмы и т.д. Но наибольшее распространение получила модель логистической регрессии [10].

В данной работе мы рассмотрим модель логистической регрессии со штрафной функцией для отбора признаков при условии наложения на модель экспертных ограничений. Отбор признаков необходим, чтобы исключить из модели нерелевантные и избыточные признаки. Это существенно упрощает ее, и улучшает качество. Экспертные ограничения же необходимы, например, в случае некорректно построенных обучающих выборок. Такой подход дает аналитику возможность ставить ограничения на определенные коэффициенты модели, исходя из экспертных знаний и здравой логики. Это крайне актуально, для недавно открывшихся банков, которые располагают пока небольшой клиентской базой и могут работать только с этим набором данных, в котором могут встречаться разного рода ошибки. Так же данные могут исказиться в периоды кризиса, когда привычный для банка портрет клиента сильно меняется, и если этот период необходимо включать в данные для построения модели, то экспертные ограничения могут существенно помочь.

Работа состоит из 4 основных разделов. В первом разделе описывается предметная область: ставится задача оценки кредитоспособности заемщика как задачи классификации, описываются основные методы ее решения, а так же приводятся способы оценки качества модели. Второй раздел посвящен модели логистической регрессии с регулируемой селективностью, рассмотрен байесовский подход, на котором построены встроенные методы отбора признаков, приводятся краткие описания самых популярных методов. В третьем разделе рассматривается вышеупомянутая модель при наличии экспертных ограничений. Четвертый раздел содержит экспериментальные исследования.

2 Анализ предметной области и постановка задачи оценки кредитоспособности заемщика

Кредитный скоринг — это система оценки кредитоспособности (кредитных рисков) лица, основанная на численных статистических методах. В банковской деятельности России наиболее распространены следующие виды скоринга:

- **Application-скоринг**. Основная задача – оценка кредитоспособности потенциального заемщика. По количеству набранных баллов, в соответствии с политикой банка, принимается решение о выдаче кредита и его условиях;
- **Fraud-скоринг**. Задача данного вида скоринга – предотвратить потенциальное мошенничество со стороны заемщика. В некоторых случаях, потенциальные заемщики злонамеренно создают ложный образ идеального заемщика, с целью хищения средств банка. Скоринг помогает выявить подобных мошенников;
- **Collection-скоринг**. Инструмент в помощь отделу по работе с неблагополучными заемщиками. Подсказывает приоритетные варианты работы с просроченной задолженностью.

Потребность в автоматической скоринговой системе возникла по следующим причинам. Во-первых, подход, основанный исключительно на экспертной оценке, считался недостаточно надежным, к тому же, человек, принимающий решения должен обладать очень высокой квалификацией и иметь большой опыт. Во-вторых, в 1960-х годах резко увеличилось количество желающих получить кредит и банки ощутили явную нехватку квалифицированных кадров для работы с потоком заемщиков.

При решении задачи кредитного скоринга строится скоринговая карта. Ее построение основывается на результатах статистической обработки данных о предыдущем опыте кредитования заемщика. История скоринговых карт берет свое начало в работе американского экономиста Д. Дюрана «Элементы риска потребительского кредитования в рассрочку». В данном исследовании была впервые применена методика классификации заемщиков на хороших и плохих, а также выявлены группы факторов, помогающих оценить кредитоспособность заемщика.

Скоринговая карта представляет собой набор характеристик заемщиков и соответствующих весов коэффициентов скоринговой функции, которые преобразуют в баллы. На рисунке 1 представлен пример простой скоринговой карты:

Показатель	Значение (диапазон значений) показателя	Скоринг-балл
Возраст	До 30 лет	30
	35 - 50 лет	35
	Старше 50 лет	28
	Среднее	22
Образование	Средне специальное	29
	Высшее	35
Состоит ли в браке	Да	25
	Нет	12
Брал ли кредит ранее	Да	41
	Нет	22
Трудовой стаж	Менее 1 года	16
	От 1 до 5 лет	19
	От 5 до 10 лет	24
	Более 10 лет	31
Наличие автомобиля	Да	49
	Нет	18
Возраст автомобиля	Менее 3-х лет	45
	От 3 до 7 лет	25
	Старше 7 лет	18

Рис 1. Пример скоринговой карты

Обычно, чем выше суммарный балл, тем более надежным является потенциальный заемщик. При построении карты так же выводится некоторое пороговое значение, и если суммарный балл заемщика ниже, то ему отказывают в выдаче кредита. Так же скоринговый балл клиента определяет его категорию, в соответствии с которой рассчитывают итоговый лимит, ставку, максимальную кредитную нагрузки и так далее.

Но нередко аналитики сталкиваются с проблемой неправильно сформированных выборок и недостатком данных для анализа. Поэтому стали рассматривать алгоритмы с экспертными ограничениями для модели. Они помогают выправить неправильные с точки зрения логики зависимости и построить хорошую модель.

2.1 Математическая постановка задачи

Задача кредитного скоринга по сути является задачей бинарной классификации. Приведем ее вероятностную постановку.

Пусть Ω - множество кандидатов на получение кредита в банке. Каждому потенциальному заемщику $\omega \in \Omega$ ставится в соответствие его признаковое описание $\boldsymbol{x} = \big(x_1(\omega), \dots, x_n(\omega)\big) \in X = R^n$, а так же значение зависимой переменной $Y = \{-1; +1\}$. Зависимая переменная описывает класс заемщика, метка +1 соответствует класса «хороших», а метка -1 – классу «плохих» клиентов, из-за которых банк понесет убытки. В задаче кредитного скоринга требуется построить такую модель, которая по признаковому

описанию **x** потенциального заемщика будет предсказывать значение зависимой переменной y.

Предполагаем, что на множестве $X \times Y$ существует какое-то вероятностное распределение и его плотность обозначим $p(\boldsymbol{x}, y)$. Из этого множества выбирается так называемая обучающая выборка $\Omega^* = \{(\boldsymbol{x}(\omega_i), y(\omega_i))\}_{i=1}^m$ – это набор случайно и независимо выбранных наблюдений, т.е. описания заемщиков. Именно на основе этой выборки будет строиться модель.

Для решения поставленной задачи обычно используется принцип максимума апостериорной вероятности. Апорстериорная вероятность того, что заемщик принадлежит «хорошему» или «плохому» классу, может быть посчитана с использованием формулы Байеса:

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|y)p(y)}{p(\boldsymbol{x})} \quad (1)$$

2.2 Обзор существующих моделей

Линейный дискриминант Фишера:

Модель использует метод максимального правдоподобия и основана на предположении о том, что функции правдоподобия классов многомерные нормальные. Основной недостаток в том, что такая гипотеза слишком сильная и не всегда выполняется.

Логистическая регрессия:

Модель напрямую оценивает вероятности принадлежностей объектов классам, используя принцип максимума правдоподобия. Строится скоринговая функция на основе логит-функции, веса функции находятся по методу наименьших квадратов. Модель основана на более слабых предположениях, чем предыдущий метод, поэтому часто работает лучше.

Логистическая регрессия является самым популярным методом в задаче кредитного скоринга из-за своей простоты и эффективности.

Метод k ближайших соседей:

Данная непараметрическая модель является одной из самых простых и понятных. Объект относится к классу, к которому относятся *k* ближайших к нему объектов выборки. Близость определяется по заранее выбранной метрике на пространстве данных об объектах.

Основной недостаток модели – необходимость полного просмотра набора объектов, когда нужно проклассифицировать новы. Кроме того, выбор правильной метрики является достаточно сложной задачей.

Деревья решений:

Модель легко интерпретируема и может строиться на основе небольших выборок. Здесь не строится линейная скоринговая функция, вместо этого берется функция одного аргумента (обычно этот аргумент – это значения определенного признака) и на ее основе объекты последовательно разделяются так, чтобы группы максимально отличались по уровню риска. Дерево строится до тех пор, пока новое разделение не приведет к статистически значимому различию в величине кредитного риска. Каждый лист построенного дерева в итоге соответствует классу.

После этого множество деревьев, построенных различными разделяющими функциями, тестируются на выборке и выбирается лучшее.

Генетические алгоритмы:

Эта модель не накладывает стандартных ограничений на целевую функцию, таких как гладкость, выпуклость и так далее. В рассматриваемой задаче генетические алгоритмы генерируют начальное множество скоринговых функций, после чего к ним применяются «мутации», «скрещивания», отбрасывание непригодных функций.

Нейронные сети:

Данная модель берет за основу принцип работы биологических нейронных сетей и их строение. Нейронные сети способны моделировать сложные нелинейные зависимости, хотя для задачи кредитного скоринга это не всегда необходимо и чаще применяется в скоринге компаний. Модели имеют достаточно сложную топологию (множество слоев разного типа, различные функции активации) и веса связей, которые получаются в результате обучения модели, не имеют никакой интерпретации для задачи кредитного скоринга. Значит, нет возможности объяснить полученное предсказание и провести анализ значимости факторов.

2.3 Предобработка данных

Существует несколько этапов предварительной обработки данных для модели. Перечислим основные из них:

Работа с пропусками и выбросами в данных:

Для построения модели требуются полные данные без экстремальных значений. При работе с пропусками следует понимать их природу: случайно ли их появление или нет. Если пропуски случайны, т.е. появились, например, за счет неверного заполнения

анкеты операционистом, тогда допустимо удаление данных, заполнение пропусков средними или наиболее вероятными значениями. Если же пропуски имеют неслучаный характер, то удаление данных из выборки запрещено и следует применять различные статистические методы, которые на основе значений других переменных будут определять значение для конкретного пропуска.

То же самое касается экстремальных значений. Здесь распространённой практикой является замена на максимальное или среднее значение.

Квантование данных:

В модель достаточно сложно встроить непрерывные переменные или категориальные переменные с большим количеством возможных значений, поэтому перед построением моделей часто производят квантование или категоризацию. Непрерывная переменная делится на несколько групп или конечных классов на основе анализа значимости конкретной категории и переменной в целом после такого разбиения. С категориальными переменными проводят тот же анализ, но в процессе объединения категорий в различные группы.

Категоризация дает возможность смоделировать нелинейные зависимости в линейной модели.

Первичный отбор признаков:

В задаче банковского скоринга обычно на начальных этапах имеется достаточно большое количество признаков, т.е. объединяются данные из анкеты заемщика, данные по его кредитной и депозитной истории, данные полученные из внешних источников и так далее. Поэтому достаточно часто перед построением модели обычно проводят начальный отсев признаков. Распространённым методом считается:

- оценка характера связи значения переменной с бинарной выходной переменной:

$$WoE_i = \ln\left(\frac{d_i^{(1)}}{d_i^{(2)}}\right)$$

где $d_i^{(1)}$ и $d_i^{(2)}$- относительные частоты «плохих» и «хороших» кредитов соответственно в i-ой группе категорированной переменной.

- оценка значимости переменной для бинарной классификации с помощью информационного индекса:

$$IV = \sum_{i=1}^{K} \left\{\left(d_i^{(1)} - d_i^{(2)}\right) WoE_i\right\}$$

Сэмплинг:

Это процесс отбора из исходной совокупности данных выборки, представляющей интерес для анализа. По итогу такого отбора выборка должна быть максимально репрезентативной. Сэмплинг бывает случайным, равномерным случайным и статифицированным. В итоге выборка должна делиться на тестовую и обучающую. Иногда выделяется отдельная часть данных, которая не используется для анализа.

Так же сэмплинг применяется для решения проблемы «редкого класса», т.е. при ситуации, когда присутствует ярко выраженная несбалансированность классов. Это может повлечь за собой следующую проблему: классификатор будет склонен будет с большей вероятностью относить новые примеры к классу с наибольшим количеством примеров, т.е. к мажоритарному. Проблема может решаться модификацией алгоритма построения классификатора или же специальными методами сэмплинга, например отбором со смещением (удалением мажоритарного класса).

2.4 Оценка качества модели

После построения модели необходимо оценить ее качество. Перечислим основные методы, применяемые в задаче кредитного скоринга.

Скользящий контроль:

Это техника оценки обобщающей способности модели является одной из самых классической. Выборка делится на обучающую и тестовую, модель строится на обучающей, а оценивается на тестовой. Среднее арифметическое ошибки по всем разбиениям называется оценкой скользящего контроля.

Выборка может разбиваться случайно, а может на k блоков примерно разной длины. В последнем случае метод называется k-fold cross validation, тогда один блок используется для контроля, а все остальные для обучения.

Таблица классификации:

Введем два важных понятия:

- **Ошибка первого рода** состоит в том, что будет отвергнута нулевая гипотеза, хотя на самом деле она верна, т.е. модель принимает «хорошего» заемщика за «плохого».
- **Ошибка второго рода** состоит в том, что будет принята нулевая гипотеза, хотя в действительности верна конкурирующая, в данном случае модель принимает «плохого» заемщика за «хорошего».

Для задачи кредитного скоринга соотношение ошибок первого и второго рода в модели осуществляются с помощью таблицы классификации и ROC-анализа.

Построение таблицы классификации позволяет оценить дискриминирующую способность модели. В ячейках этой таблицы приводится количество фактических и предсказанных значений зависимой переменной по каждому классу, на основе которых вычисляется процент корректных предсказаний по каждому классу (см. табл. 1).

Прогноз того, является ли клиент кредитоспособным	Фактическая кредитоспособность		Процент корректных предсказаний
	Да	Нет	
Да	710	250	73,96%
Нет	120	520	67,57%
Итого	76,88%		

Таблица 1. Пример таблицы классификации

Анализ ROC-кривой:

ROC-кривая показывает соотношение ошибок первого и второго рода и помогает выбрать порог вероятности, которая должна отделять один класс от другого. Аналитик может менять порог отсечения и управлять соотношением ошибок первого и второго рода.

В рассматриваемой задаче ROC-кривая строится по значениям спрогнозированных вероятностей успешного погашения кредита. По одной оси откладывается количество верно классифицированных положительных примеров (чувствительности), по другой - количество неверно классифицированных отрицательных примеров (единица минус специфичность). Идеальная модель обладает 100% чувствительностью и специфичностью. Если модель обладает высокой специфичностью, то банк чаще отказывает в выдаче кредита, и при высокой чувствительности ведет более рискованную политику выдачи кредитов.

В кредитном скоринге цена ошибки второго рода выше, т.к. связана с серьезными потерями для банка из-за невозвращенных кредитов, поэтому модель должна быть больше

ориентирована на классификацию «плохих» заемщиков. В соответствии с этим должен выбираться оптимальное значение порога отсечения.

Основным показателем для анализа ROC-кривой является показатель AUC (от англ. Area Under Curve), равный площади под графиком, измеряющийся в диапазоне от 0. 5 (нет разделения) до 1 (идеальное разделение). Данный показатель используется для сравнения моделей между собой.

Считается, что значение площади:

- От 0.5 до 0.6 соответствует неудовлетворительному качеству модели
- От 0.6 до 0.7 соответствует среднему качеству модели
- От 0.7 до 0.8 соответствует хорошему качеству модели
- От 0.8 до 0.9 соответствует очень хорошему качеству модели
- От 0.9 до 1 соответствует отличному качеству модели.

Вместо значения AUC часто используют индекс Джини. Этот показатель переводит значение площади под кривой в диапазон от 0 до 1, и чем он больше, тем выше прогностическая способность модели. Рассчитывается индекс Джини по формуле:

$$D = 2 * AUC + 1$$

Статистика Колмогорова-Смирнова:

Кроме индекса AUC ROC-кривая инкапсулирует в себе еще одну важную метрику оценки качества – статистику Колмагорова-Смирнова (статистика KS). Эта величина равно максимальной разнице между долей инстинноположительных и ложноположительных объектов. Метрику принято измерять в процентах и ее значение равносильно максимальной длине вертикальной линии между ROC-кривой и диагональной линией бесполезного классификатора. Модель считается хорошей при статистике KS более 40%.

Так же после построения скоринговой карты по модели стоит обратить внимание на следующие диаграммы:

Распределение скорингового балла:

Распределение скорингового балла - это простой и наглядный способ визуальной оценки получившейся карты. Аналитиком строится гистограмма, где каждому столбцу присваивается число заемщиков и соответствующий им диапазон скоринговых баллов.

Такой график должен иметь холмообразную форму и быть максимально близким к кривой нормального распределения.

Распределение событий/не-событий

В задаче кредитного скоринга событием считается то, что заемщик «хороший», не-событием, соответственно, что заемщик является «плохим». Очень важно, чтобы в скоринговой карте объекты, с которыми происходит событие, и объекты, с которыми не происходит, имели разные баллы.

Строятся графики распределений событий и не-событий, на которых ставится соответствие количество счетов и скоринговые баллы. Идеальной считается карта, на которой эти распределения не пересекаются вообще, а находятся рядом. Если карта неэффективна, то графики распределения будут налагаться друг на друга, т.е. почти невозможно различить группы событий и не-событий.

3 Модель логистической регрессии с регулируемой селективностью

3.1 Задача отбора признаков

В задаче кредитного скоринга на вход подается достаточно большое количество признаков: данные по кредитной истории заемщика, его анкетные данные, депозитная история и так далее. Поэтому проблема отбора релевантных признаков в модель становится крайне актуальной. Следует исключать из анализа нерелевантные и коррелируемые между собой признаки.
Правильный отбор признаков позволяет решить следующие задачи:

- Увеличение обобщающей способности модели
- Повышение интерпретируемости модели
- Снижение сложности и времени вычисления модели

3.1.1 Обзор существующих подходов

Выделяют три основных подхода к отбору признаков [7,17]. Они отличаются способом комбинирования алгоритма отбора и алгоритма обучения модели.

- **Методы-фильтры (filter-based methods)**. Эти методы применяются, когда модель уже построена и не зависят от алгоритма обучения. Они базируются на статистическом анализе и оценивают, насколько взаимосвязан признак и целевая переменная. Если связь слаба, то признак исключается из модели.

Методы эффективны с точки зрения вычислительных ресурсов, но не учитывают влияние выбранного подмножества признаков на качество итогового классификатор

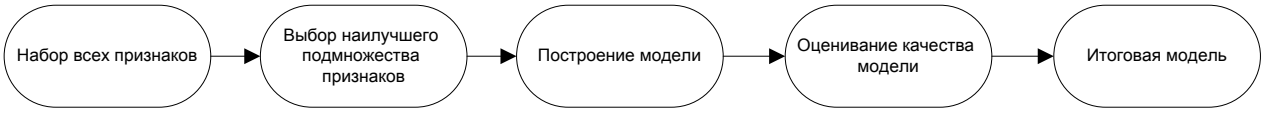

Рис 2. Отбор признаков методом-фильтром

- **Методы-обертки (wrapper-based methods)**. Методы этого класса оценивают подмножество признаков исходя их качества модели, построенной с их помощью.

Если N – это число признаков, то для нахождения оптимального набора признаков при полном переборе необходимо проверить 2^N комбинаций, т.е. в общем случае задача отбора признаков NP-трудная и для её решения требуется экспоненциальное время.

Поэтому сокращения перебора применяются различные субоптимальные методов, которые опираются на различные эвристики, позволяющие сократить время перебора, например:

- прямой отбор (Forward Selection)
- обратное исключение (Backward Elimination)
- последовательный отбор (Stepwise procedure)
- различные метаэвристики (например, генетический алгоритм, поиск с запретами и т.д.).

Как было показано в исследованиях [14], в задаче кредитного скоринга методы этого класса работают лучше, чем методы-фильтры [14]. Но они требуют многократного обучения классификатора и поэтому вычислительно они достаточно затратны.

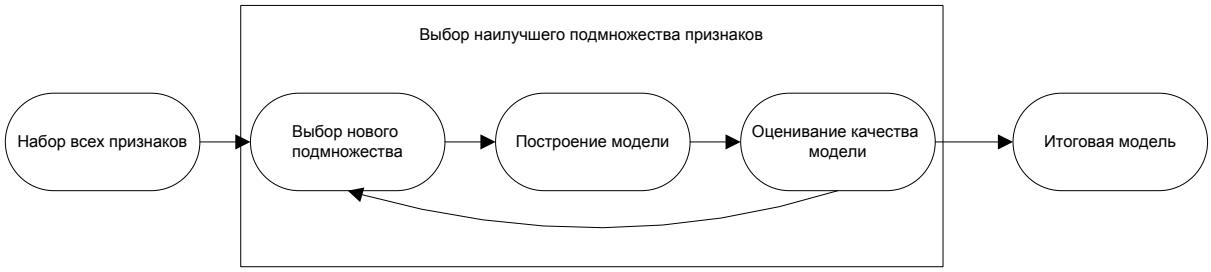

Рис 3. Отбор признаков методом-оберткой

- **Встроенные методы (embedded methods).** Встроенные методы имеют преимущество методов-фильтров и методов-оберток, т.к. алгоритм отбора признаков включен в алгоритм обучения модели. Они учитывают влияние подмножества признаков

на результат работы модели и вычислительны не так затратны, как методы обертки. Именно поэтому в последнее время они набирают популярность

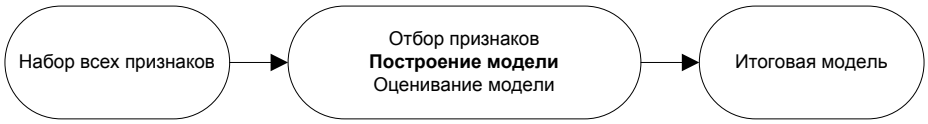

Рис 4. Встроенный отбор признаков

3.1.2 Встроенные методы отбора признаков на основе байесовского подхода

Наиболее часто используемым встроенным методом отбора признаков является регуляризация. Основная идея заключается во включении в целевую функцию слагаемого (регуляризатора), который «штрафует» коэффициенты модели, устремляя их к нулю.

При регуляризации вектор параметров **w** рассматривается как вектор случайных чисел с априорным распределением $p(\boldsymbol{w})$.

Апостериорная плотность распределения параметров ищется по формуле Байеса:

$$p(\boldsymbol{w}|\Omega^*) = \frac{p(\Omega^*|\boldsymbol{w})p(\boldsymbol{w})}{p(\Omega^*)} \tag{2}$$

Т.к. $p(\Omega^*)$ не зависит от **w**, следовательно:

$$p(\boldsymbol{w}|\Omega^*) \propto p(\Omega^*|\boldsymbol{w})p(\boldsymbol{w}) \tag{3}$$

Используя принцип максимизации апостериорной плотности, получаем точечную оценку значений вектора параметров:

$$\widehat{\boldsymbol{w}}_{MAP} = arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\,\Omega^*) = arg\max_{\boldsymbol{w}} p(\Omega^*|\boldsymbol{w})p(\boldsymbol{w}) \tag{4}$$

$$\widehat{\boldsymbol{w}}_{MAP} = arg\min_{\boldsymbol{w}}\{-\ln p(\Omega^*|\boldsymbol{w}) - \ln p(\boldsymbol{w})\} \tag{5}$$

В формуле (5) второе слагаемое является вышеупомянутым регуляризатором или, так называемой, штрафной функцией. Обычно штрафная функция включается в модель с коэффициентом, с помощью которого можно контролировать количество отбираемых в модель признаков.

Перечислим самые распространенные методы регуляризации:

- **Гребневая регрессия (ridge).** В данном методе в качестве априорного распределения выбирается нормальное распределение. Штрафная функция будет выглядеть следующим образом:

$$penalty(\boldsymbol{w}) = \lambda \sum_{i=1}^{n} w_i^2 \quad (6)$$

Метод сжимает коэффициенты. Количество признаков не меняется, но понижается эффективная размерность задачи.

- **Метод Lasso**. В данном методе в качестве априорного распределения выбирается закон Лапласа. Штрафная функция примет следующий вид:

$$penalty(\boldsymbol{w}) = \lambda \sum_{i=1}^{n} |w_i| \quad (7)$$

Данный метод производит отбор признаков, но при этом если признаки сильно коррелированны, то отберется только один из них, что является недостатком.

- **Метод Elastic Net.** Данный метод комбинирует два предыдущих:

$$penalty(\boldsymbol{w}) = (1 - \lambda) \sum_{i=1}^{n} w_i^2 + \lambda \sum_{i=1}^{n} |w_i| \quad (8)$$

Основной целью его создания было желание преодолеть неспособность метода Lasso отбирать коррелируемые признаки в модель.

3.1.3 Свойства оценок параметров моделей со встроенными методами отбора признаков

Для оценивания качества встроенных методов отбора признаков в литературе [5] выделяют следующие критерии:

- **Асимптотическая несмещенность.** С точки зрения байесовского модели данный критерий означает то, что для отобранных в модель признаков априорное распределение коэффициентов регрессии является равномерным.
- **Непрерывность.** Критерий характеризует то, что отбор признаков осуществляется для всех значений структурного параметра и одновременно для всех признаков. Т.е. маленькое изменение в данных не будет приводить к резкому изменению отобранных признаков.
- **Селективность**. Критерий характеризует способность исключать из модели незначащие признаки.

- **Оракульные свойства**. Критерий выражает вероятность совпадения множества отобранных алгоритмом признаков с множеством признаков, действительно присутствующих в модели, а также оценивающие квадратичную сходимость абсолютных значениях регрессионных коэффициентов к их истинным значениям.
- **Способность к отбору коррелированных признаков**. Это свойство характеризует способность оставлять в итоговой модели значащие признаки даже в случае наличия корреляции между ними.

Как видно из приведенной ниже таблицы, ни один из основных методов не обладает всеми свойствами.

	Ridge	**Lasso**	**Elastic Net**
Селективность	Нет	Да	Да
Непрерывность	Нет	Нет	Нет
Асимптотическая несмещенность	Нет	Нет	Нет
Оракульные свойства	Да	Да	Да
Способность к отбору коррелированных признаков	Да	Нет	Да

Таблица 2. Преимущества и недостатки существующих методов отбора признаков

3.2 Разработка модели

При решении задачи бинарной классификации условное распределение зависимой переменной представляет собой распределение Бернулли:

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = Ber(y|sigm(\boldsymbol{w}^T\boldsymbol{x})) \quad (9)$$

где $sigm(\boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{w}^T x}}$ – логистическая функция (сигмоида), а $\boldsymbol{w}^T\boldsymbol{x} = \sum_{i=1}^{n} w_i x_i$- линейная разделяющая гиперплоскость.

Тогда вероятности того, что заемщик принадлежит к классам «плохих» и «хороших» равны соответственно:

$$P(y = +1|\boldsymbol{x}, \boldsymbol{w}) = sigm(\boldsymbol{w}^T\boldsymbol{x}) \quad (10)$$

$$P(y = -1|\boldsymbol{x}, \boldsymbol{w}) = 1 - sigm(\boldsymbol{w}^T\boldsymbol{x}) \quad (11)$$

Можно записать это одним выражением:

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = sigm(y\boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1 + e^{-y\boldsymbol{w}^T\boldsymbol{x}}} \quad (12)$$

Мы предполагали, что наблюдения в обучающей выборке независимы, поэтому функция правдоподобия будет выглядеть следующим образом:

$$L(\boldsymbol{w}|\Omega^*) = \prod_{i=1}^{m} p(y_i|x_i, \boldsymbol{w}) = \prod_{i=1}^{m} sigm\left(y_i \sum_{j=1}^{n} w_j x_{ij}\right) \quad (13)$$

Используя принцип максимума правдоподобия, получаем оценку вектора параметров:

$$\widehat{\boldsymbol{w}}_{ML} = arg\max_{\boldsymbol{w}}\left\{\prod_{i=1}^{m} sigm\left(y_i \sum_{j=1}^{n} w_j x_{ij}\right)\right\} \quad (14)$$

Прологарифмируем функцию правдоподобия (14) и будем решать задачу минимизации:

$$\widehat{\boldsymbol{w}}_{ML} = arg\min_{\boldsymbol{w}}\left\{-\sum_{i=1}^{m} \ln\left(sigm\left(y_i \sum_{j=1}^{n} w_j x_{ij}\right)\right)\right\} \quad (15)$$

В соответствии с (15) оценивают коэффициенты в классической нерегуляризованной логистической регрессии.

Предположим, что априорной плотностью распределения параметра является нормальное распределение с нулевым матожиданием и дисперсией **r**. В модели дисперсия будет являться случайной величиной.

Тогда совместное распределение вектора параметров будет иметь вид:

$$p(\boldsymbol{w}|\boldsymbol{r}) \propto \prod_{i=1}^{n}\left[\left(\frac{1}{r_i}\right)^{\frac{1}{2}} \exp(-\sum_{i=1}^{n}\frac{w_i^2}{2r_i})\right] \quad (16)$$

Параметры с малым значением r_i могут быть удалены из модели, а остальные параметры будут называться релевантными. Здесь r_i – является гиперпараметром модели.

Попробуем величины, обратные дисперсиям, использовать в качестве штрафных функций. Тогда предполагаем, что априорная плотность распределения величин обратных дисперсиям является гамма-распределением:

$$p\left(\frac{1}{\boldsymbol{r}}|\alpha, \beta\right) \propto \prod_{i=1}^{n}\left[\left(\frac{1}{r_i}\right)^{\alpha-1} \exp(-\frac{\beta}{r_i})\right] \quad (17)$$

Из (17) видно, что обратная дисперсия зависит от двух параметров гамма-распределения α, β. Для облегчения процесса подбора параметров предположим, что они являются функциями от одного и того же параметра μ.

Для случайной величины, имеющей гамма-распределение, известно:

- $E\left(\frac{1}{r_i}\right) = \frac{\alpha}{\beta}$ – математическое ожидание
- $Var\left(\frac{1}{r_i}\right) = \frac{\alpha}{\beta^2}$ – дисперсия

Рассмотрим отношение $\frac{\sqrt{Var\left(\frac{1}{r_i}\right)}}{E\left(\frac{1}{r_i}\right)} = \frac{1}{\sqrt{\alpha}}$:

- Если $\frac{\sqrt{Var\left(\frac{1}{r_i}\right)}}{E\left(\frac{1}{r_i}\right)} \to 0$, то значит все распределения дисперсий r_i сконцентрированы возле математического ожидания. Тогда можно сказать, что оцененные дисперсии практически фиксированы и равны единице при $\alpha \cong \beta$.
- Если $\frac{\sqrt{Var\left(\frac{1}{r_i}\right)}}{E\left(\frac{1}{r_i}\right)} \to 1$, то априорные распределения становятся практически равномерными.

При $r_i \to 0$: $\ln r_i \to -\infty$ и критерию выгодно уменьшать все дисперсии. Но в этом случае невозможно выполнить ограничения, предписывающие достаточно хорошо приближать обучающую совокупность. Из-за этого противоречия критерий проявляет ярко выраженную склонность к чрезмерной селективности отбора признаков, подавляя большинство из них, в том числе и релевантные.

Получается, что необходимо выполнение следующих требований:

$$\mu \to 0 \Rightarrow \begin{cases} E\left(\frac{1}{r_i}\right) \to 1 \\ Var\left(\frac{1}{r_i}\right) \to 0 \\ \frac{\sqrt{Var\left(\frac{1}{r_i}\right)}}{E\left(\frac{1}{r_i}\right)} \to 0 \end{cases} \quad \mu \to \infty \Rightarrow \begin{cases} E\left(\frac{1}{r_i}\right) \to \infty \\ Var\left(\frac{1}{r_i}\right) \to \infty \\ \frac{\sqrt{Var\left(\frac{1}{r_i}\right)}}{E\left(\frac{1}{r_i}\right)} \to 1 \end{cases} \tag{18}$$

Одним из наборов функций, удовлетворяющих требованиям, является:

$$\alpha = 1 + \frac{1}{2\mu} \text{ и } \beta = \frac{1}{2\mu} \tag{19}$$

С учетом всех предположений об априорных распределениях вектора параметров и гиперпараметров получаем следующую оценку вектора параметров:

$$\widehat{\boldsymbol{w}}_{MAP} = arg\max_{\boldsymbol{w}}\{p(y|\boldsymbol{w},\boldsymbol{x})p(\boldsymbol{w}|\boldsymbol{r})p(\boldsymbol{r}|\alpha,\beta)\} \tag{20}$$

$$\widehat{\boldsymbol{w}}_{MAP} = arg\min_{\boldsymbol{w}}\{-\ln p(y|\boldsymbol{w},\boldsymbol{x}) - \ln p(\boldsymbol{w}|\boldsymbol{r}) - \ln p(\boldsymbol{r}|\alpha,\beta)\} \tag{21}$$

Тогда получаем следующий критерий обучения:

$$J(\boldsymbol{w},\boldsymbol{r}) = -\sum_{i=1}^{m} \ln\left(sigm\left(y_i \sum_{j=1}^{n} w_j x_{ij}\right)\right) + \frac{1}{2}\sum_{i=1}^{n} \frac{w_i^2}{r_i} + \left(\alpha - \frac{1}{2}\right)\sum_{i=1}^{n} \ln r_i + \beta \sum_{i=1}^{n} \frac{1}{r_i} \to \min_{\boldsymbol{w},\boldsymbol{r}} \quad (22)$$

Подставим в критерий выбранные функции для α и β:

$$J(\boldsymbol{w},\boldsymbol{r}) = -\sum_{i=1}^{m} \ln\left(sigm\left(y_i \sum_{j=1}^{n} w_j x_{ij}\right)\right) + \frac{1}{2}\sum_{i=1}^{n} \frac{w_i^2}{r_i} + \left(1 + \frac{1}{\mu}\right)\sum_{i=1}^{n} \ln r_i + \frac{1}{2\mu} \sum_{i=1}^{n} \frac{1}{r_i} \to \min_{\boldsymbol{w},\boldsymbol{r}} \quad (23)$$

Этот критерий будем называть моделью логистической регрессии с регулируемой селективностью. Он удовлетворяет всем свойствам, приведенным в пункте 2.1.3.

4 Разработка модели логистической регрессии с регулируемой селективностью при наличии экспертных ограничений

Многие российские банки встречаются с проблемами недостатка данных для построения скоринговых моделей. Эта проблема только усугубляется, если выборки еще и некорректны. Для решения предлагается задавать дополнительные экспертно-интерпретируемые ограничения на коэффициенты модели.

Рассмотрим дополнительные ограничения на коэффициенты модели вида:

- $sign(w_i) > 0$
- $w_{i+1} \geq w_i$
- $w_i \leq t$
- $\|w_i\| \leq t$

Такие ограничения устанавливаются аналитиками, решающими конкретную задачу и называются предметно-экспертными. Такие ограничения в некоторых случаях могут повысить обобщающую способность или качество модели. Особенно они полезны в случае некорректной или несбалансированной обучающей выборки.

Приведем примеры, когда предметно-экспертные ограничения могут повысить качество модели:

- Может быть экспертно установлено, что чем больше значение какого-то признака, тем выше риски. Например, таким признаком может быть размер

запрашиваемой суммы кредита. Тогда признак должен выйти в модель с положительным коэффициентом и имеет смысл ввести следующее ограничение:

$$w_{required\_limit} \geq 0$$

Обратная ситуация со сроком кредита. Обычно, чем меньше запрашиваемый срок, тем больше риски и ограничение должно быть следующим:

$$w_{required\_term} < 0$$

- Крайне часто в анализе встречаются признаки, которые надо квантовать, т.е. разбивать на интервалы значений. Может быть выяснено, что интервалы по разному связаны с риском. Например, если в $i+1$-ом интервале вероятность дефолта больше, чем в i-ом, то имеет смысл ввести следующее ограничение:

$$w_{i+1} \geq w_i$$

- Ограничение на норму подвектора вводят для уменьшения эффекта переобучения. Это может помогать в случае, когда есть бинарные признаки, принимающие значение 1 только на объектах целевого класса и каждый из них покрывает только малую часть целевого класса. Обычно модель сначала обучают без ограничений, находят норму вектора коэффициентов и потом в качестве ограничения берут какую-то ее долю.

Накладывая экспертные ограничения на модель логистической регрессии с регулируемой селективностью, получаем следующую задачу минимизации.

$$\min_{\boldsymbol{w}\in\mathbb{R}^n}\left(-\sum_{i=1}^{m}\ln\left(sigm\left(y_i\sum_{j=1}^{n}w_j x_{ij}\right)\right)+\frac{1}{2}\sum_{i=1}^{n}\frac{w_i^2}{r_i}+\left(1+\frac{1}{\mu}\right)\sum_{i=1}^{n}\ln r_i + \frac{1}{2\mu}\sum_{i=1}^{n}\frac{1}{r_i}\right) \tag{24}$$

при линейных ограничениях типа:

$\varphi(\boldsymbol{w}) \leq 0$, где $\varphi(\boldsymbol{w})$ – выпуклая функция.

Для ее решения предлагается использовать метод штрафных функций.

4.1 Метод штрафных функций

Рассмотрим задачу минимизации с ограничениями:

$$f^* = \min_{x\in R^n} f(x)$$
$$\varphi_i(x) \leq 0, i = 1, \dots, m \tag{25}$$

$$\varphi_i(x) = 0, i = m+1, \dots, l$$

Пусть ограничения задают множество $G \in R^n$. Определим индикаторную функцию множества *G* следующим образом:

$$\delta(\boldsymbol{x}|G) = \begin{cases} 0, & x \in G \\ +\infty, & x \notin G \end{cases} \quad (26)$$

Допустим: $F(x) = f(x) + \delta(\boldsymbol{x}|G)$.

Тогда задача (25) эквивалентна следующей задаче безусловной минимизации:

$$\min_{\boldsymbol{x} \in R^n} F(x) \quad (27)$$

Пусть: $\delta(\boldsymbol{x}|G) = \lim_{k\to\infty} \delta_k(\boldsymbol{x}|G)$.

$\delta_k(\boldsymbol{x}|G)$ назовем штрафными функциями и вместо задачи (27) будем решать следующую задачу:

$$\min_{\boldsymbol{x} \in R^n} F_k(x) = \min_{\boldsymbol{x} \in R^n} f(x) + \delta_k(\boldsymbol{x}|G) \quad (28)$$

$$x_k^* = arg \min_{\boldsymbol{x} \in R^n} F_k(x) \quad (29)$$

Теорема 1. Пусть *f(x)* непрерывна на R^n, так же пусть заданы штрафные функции $\delta_k(\boldsymbol{x}|G)$, такие что:

- $\delta_k(\boldsymbol{x}|G) = 0,\ x \in G$
- $\delta_k(\boldsymbol{x}|G) > 0,\ x \notin G$
- $\delta_{k+1}(\boldsymbol{x}|G) > \delta_k(\boldsymbol{x}|G),\ x \notin G$
- $\lim_{k\to\infty} \delta_k(\boldsymbol{x}|G) = \delta(\boldsymbol{x}|G)$

$x_k^* = arg \min_{\boldsymbol{x} \in R^n} F_k(x)$, тогда существует предельная точка x^*: $x^* = \lim\limits_{k\to\infty} x_k^* = f(x^*) = f^*$

Приведем пример применения метода.

Допустим, требуется решить следующую задачу:

$$\begin{gathered} \min f(x_1, \dots, x_n) \\ \text{при } g_i(x_1, \dots, x_n) \geq 0\ i = 1,2 \dots m \end{gathered} \quad (30)$$

Тогда введем функцию:

$$F(x_1, \dots, x_n) = f(x_1, \dots, x_n) + P(x_1, \dots, x_n) \quad (31)$$

Здесь *P* – это штрафная функция, которую можно выбрать в виде:

$$P(x_1, \dots, x_n) = c \sum_{i=1}^{m} g_i^2(x_1, \dots, x_n)$$

либо

$$P(x_1, \dots, x_n) = c \sum_{i=1}^{m} 1/g_i(x_1, \dots, x_n) \quad (32)$$

где c – положительный параметр

Возьмем конкретный вид функции:

$f(x_1, x_2) = (x_1 - 1)^2 + (x_2 - 5)^3$ при $x_1 - x_2 + 10 = 0$.

Тогда минимизировать следует следующую функцию:

$F = (x_1 - 1)^2 + (x_2 - 5)^3 + c(x_1 - x_2 + 10)^2$

4.2 Модель логистической регрессии с регулируемой селективностью при наличии экспертных ограничений

4.2.1 Описание модели

Вместо задачи (24) предлагается решать следующую задачу:

$$\min_{\boldsymbol{w} \in \mathbb{R}^n} \left(-\sum_{i=1}^{m} \ln\left(sigm\left(y_i \sum_{j=1}^{n} w_j x_{ij} \right) \right) + \frac{1}{2} \sum_{i=1}^{n} \frac{w_i^2}{r_i} + \left(1 + \frac{1}{\mu}\right) \sum_{i=1}^{n} \ln r_i + \frac{1}{2\mu} \sum_{i=1}^{n} \frac{1}{r_i} + \delta_k(\boldsymbol{w}) \right) \quad (33)$$

где $\delta_k(\boldsymbol{w}) = c_k \left(\varphi(\boldsymbol{w}) \right)_+^2$, $k = 1,2, \dots$

$c_{k+1} > c_k$ – положительные коэффициенты

Обозначим:

- $L(\boldsymbol{w}) = -\sum_{i=1}^{m} \ln\left(sigm\left(y_i \sum_{j=1}^{n} w_j x_{ij} \right) \right)$
- $P(\boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{n} \frac{w_i^2}{r_i} + \left(1 + \frac{1}{\mu}\right) \sum_{i=1}^{n} \ln r_i + \frac{1}{2\mu} \sum_{i=1}^{n} \frac{1}{r_i}$.

Функции $L(\boldsymbol{w})$ и $P(\boldsymbol{w})$ являются непрерывными и выпуклыми. Функции $\delta_k(\boldsymbol{w}) = c_k \left(\varphi(\boldsymbol{w}) \right)_+^2$ удовлетворяют свойствам внешних штрафных функций, если $c_{k+1} > c_k > 0$. Тогда по Теореме 1 получаем, что вместо задачи (24) можно решать задачу (33).

4.2.2 Свойства оценки параметров модели

- **Асимптотическая несмещенность**

Данное свойство с точки зрения байесовской модели выражается в том, что для отобранных признаков априорное распределения коэффициентов является равномерным. В [20] Фан и Ли показали, что достаточное условие несмещенности можно записать следующим образом:

$$p(|w_i|)'_{w_i} \to 0 \text{ при } |w_i| \to \infty$$

Видно, что для любого значения параметра $\mu > 0$ данное требование удовлетворяется, т.к. мы используем линейные функции $\varphi(w_i)$:

$$\begin{aligned} \lim_{w_i \to \infty} \left[p(|w_i|)'_{w_i}\right] &= \lim_{w_i \to \infty} \left[\frac{\mu + 1}{\mu w_i^2 + 1} w_i + 2c_k \varphi(w_i)\varphi'(w_i)\right] \\ &= \lim_{w_i \to \infty} \left[\frac{\mu + 1}{\mu w_i^2 + 1} w_i + 2c_k \varphi(w_i)\right] \end{aligned} \tag{34}$$

- **Непрерывность**

Фан и Ли в [20] показали, что необходимое и достаточное условие непрерывности является следующее:

$$arg \min_{w_i}\{|w_i| + p'(|w_i|)\} = 0$$

Видно, что условие выполняется для вышеописанного критерия.

- **Селективность**

Вышеописанный критерий строился таким образом, что незначащим признакам присваиваются очень маленькие веса. Фактически, это приводит к исключению признаков из модели в соответствии с выбранным порогом значимости, но жесткого исключения не происходит.

- **Оракульные свойства**

Обозначим вектор реальных параметров w^0. Введем обозначения:

$$\begin{aligned} a_m &= \max\{p'_\mu(|w_i^0|): w_i^0 \neq 0\} \\ b_m &= \max\{p''_\mu(|w_i^0|): w_i^0 \neq 0\} \end{aligned} \tag{35}$$

В [20] Фан и Ли показали, что оценка обладает оракульными свойствами, если удовлетворены условия:

$$b_m \to 0, \qquad a_m\sqrt{m} \to 0 \text{ при } m \to \infty \tag{36}$$

$p_\mu(w_i) = \left(\frac{1+\mu}{2\mu}\right)\ln(\mu w_i^2 + 1) + c_k\left(\varphi\left(w_i\right)\right)_+^2$, тогда:

$$p'_\mu(w_i) = \frac{(1+\mu)|w_i|}{\mu w_i^2+1} + 2c_k\varphi\left(w_i\right)\varphi'\left(w_i\right) = \frac{(1+\mu)|w_i|}{\mu w_i^2+1} + 2c_k\varphi\left(w_i\right)$$
$$p_\mu''^{(w_i)} = \frac{(1+\mu)(1-\mu w_i^2)}{(\mu w_i^2+1)^2} + 2c_k\left(\varphi'\left(w_i\right)^2 + \varphi\left(w_i\right)\varphi''\left(w_i\right)\right) = \frac{(1+\mu)(1-\mu w_i^2)}{(\mu w_i^2+1)^2} + 2c_k \quad (37)$$

Мы рассматриваем линейные экспертные ограничения, поэтому видно, что при подходящем μ и $w_i \to \infty$ полученные оценки удовлетворяют оракульным свойствам.

- **Способность к отбору коррелированных признаков**

Пусть $h(\boldsymbol{w}) = \sum_{i=1}^n \ln(\mu w_i^2 + 1) \propto \sum_{i=1}^n p(w_i)$. Тогда матрица Гесса для регуляризирующего члена будет иметь вид:

$$\nabla^2 h = \left[\frac{\partial^2 h}{\partial w_i \partial w_j}\right]$$
$$\text{где } \frac{\partial^2 h}{\partial w_i \partial w_j} = \begin{cases} \frac{1-\mu w_i^2}{(\mu w_i^2+1)^2}, i = j \\ 0 \qquad\quad , i \neq j \end{cases} \quad (38)$$

Если $\forall w_i \in [-1/\mu\,, 1/\mu] \implies \frac{\partial^2 h}{\partial w_i^2} \geq 0$, тогда $\nabla^2 h$ - неотрицательно определена, поэтому $h(\boldsymbol{w})$ - выпуклая функция на $[-1/\mu\,, 1/\mu]$. Чем меньше μ, тем больше область, в которой метод обладает способностью к отбору коррелированных регрессоров.

При использовании логистической регрессии предположим:

$$J(\mu, \boldsymbol{w}) = -\sum_{i=1}^m \left(\boldsymbol{y} - \ln\left(sigm(\boldsymbol{X}\boldsymbol{w})\right)\right) + \left(\frac{1+\mu}{2\mu}\right)\ln(\mu w_i^2 + 1) + c_k\left(\varphi\left(w_i\right)\right)_+^2$$
$$\widehat{\boldsymbol{w}} = arg\min_{\boldsymbol{w}} J(\mu, \boldsymbol{w})$$
$$\sum_{i=1}^n x_{ij} = 0, \sum_{i=1}^n {x_{ij}}^2 = 1 \quad (39)$$
$$\rho = correlation(x_i, x_j)$$

Теорема 2. $\forall \widehat{w}_i \in [-1/\mu\,, 1/\mu]\ i = 1, \ldots, n\,, \left|\widehat{w}_i - \widehat{w}_j\right| \to 0$ когда $\rho \to 1$

Доказательство:

$\sum_{i=1}^n x_{ij} = 0, \sum_{i=1}^n {x_{ij}}^2 = 1 \implies \rho = {x_i}^T x_j$ и $\left|x_i - x_j\right|^2 = 2(1-\rho)$

$J(\boldsymbol{w}, \mu)$ – выпуклая на $[-1/\mu\,, 1/\mu]$ функция.

$\widehat{\boldsymbol{w}} \in [-1/\mu\,, 1/\mu]$ поэтому $J(\widehat{\boldsymbol{w}}, \mu) \leq J(\boldsymbol{w} = 0, \mu)$

Значит: $|\hat{r}(\mu)|^2 + \left(\frac{1+\mu}{2\mu}\right)\ln(\mu w_i^2 + 1) + c_k\left(\varphi\left(w_i\right)\right)_+^2 \leq |y|^2$,

где $\hat{r}(\mu) = \boldsymbol{y} - sigm(\boldsymbol{X}\widehat{\boldsymbol{w}}) \implies |\hat{r}(\mu)| \leq |\boldsymbol{y}|$.

Кроме того, $\frac{\partial J(\boldsymbol{w},\mu)}{\partial w_{\boldsymbol{k}}} = 0$ при $\boldsymbol{w} = \widehat{\boldsymbol{w}}$**.** Тогда:

$$-x_i{}^T\big(1 - sigm(\boldsymbol{X}\widehat{\boldsymbol{w}})\big) + \Big(1 + \frac{1}{\mu}\Big)\frac{\mu\widehat{w}_i}{\mu\widehat{w}_i{}^2 + 1} = 0$$

$$-x_j{}^T\big(1 - sigm(\boldsymbol{X}\widehat{\boldsymbol{w}})\big) + \Big(1 + \frac{1}{\mu}\Big)\frac{\mu\widehat{w}_j}{\mu\widehat{w}_j{}^2 + 1} = 0$$

$$\Longrightarrow \frac{\widehat{w}_i}{\mu\widehat{w}_i{}^2 + 1} - \frac{\widehat{w}_j}{\mu\widehat{w}_j{}^2 + 1} = \frac{1}{1+\mu}\big(x_i{}^T - x_j{}^T\big)\hat{r}(\mu)$$

$$\left|\frac{\widehat{w}_i}{\mu\widehat{w}_i{}^2 + 1} - \frac{\widehat{w}_j}{\mu\widehat{w}_j{}^2 + 1}\right| = \left|\frac{1}{1+\mu}\big(x_i{}^T - x_j{}^T\big)\hat{r}(\mu)\right| \le \frac{1}{1+\mu}\left|x_i{}^T - x_j{}^T\right||\boldsymbol{y}| = \frac{2(1-\rho)}{1+\mu}$$

$\frac{\widehat{w}_i}{\mu\widehat{w}_i{}^2+1}$ и $\frac{\widehat{w}_j}{\mu\widehat{w}_j{}^2+1}$ – это значения функции $k(t) = \frac{t}{\mu t+1}$. Эта функция возрастает на $[-1/\mu\,, 1/\mu]$. Тогда $\big|k(w_i) - k\big(w_j\big)\big| \to 0 \Leftrightarrow \big|w_i - w_j\big| \to 0$.

Тогда при $\rho \to 1 \Longrightarrow \big|\widehat{w}_i - \widehat{w}_j\big| \to 0$.

Теорема доказана.

Значит, рассматриваемый критерий обладает способностью отбирать коррелированные регрессоры.

4.3 Настройка параметров модели

4.3.1 Минимизация критерия

Для минимизации выведенного критерия будем использовать покоординатный спуск. Для этого найдем градиент функции (33) по **w.**

Вспомним, что производная сигмоиды выглядит следующим образом:

$$sigm'(x) = sigm(x)(1 - sigm(x)) \tag{40}$$

Тогда получим:

$$\frac{\partial}{\partial w_k}J(\boldsymbol{w},\boldsymbol{r}) =$$

$$-\sum_{i=1}^{m}\big(1 - sigm(y_i w^T x_i)\big)y_i x_{ik} + \frac{w_k}{r_k} + 2c_k\,\varphi(w_k)\varphi'(w_k) = -\sum_{i=1}^{m}\big(1 - \tag{41}$$

$$sigm(y_i w^T x_i)\big)y_i x_{ik} + \frac{w_k}{r_k} + 2c_k\,\varphi(w_k)$$

$$\frac{\partial}{\partial w_k w_l}J(\boldsymbol{w},\boldsymbol{r}) =$$

$$-\sum_{i=1}^{m} sigm(y_i w^T x_i)\big(1 - sigm(y_i w^T x_i)\big)x_{il}x_{ik} + \frac{1}{r_k} + 2c_k\big(\big(\varphi'(w_k)\big)^2 + \tag{42}$$

$$\varphi(w_k)\varphi''(w_k)\big) = -\sum_{i=1}^{m} sigm(y_i w^T x_i)\big(1 - sigm(y_i w^T x_i)\big)x_{il}x_{ik} + \frac{1}{r_k} + 2c_k$$

$$\frac{\partial}{\partial r_k}J(\boldsymbol{w},\boldsymbol{r}) = -\frac{1}{2}\frac{w_k{}^2}{r_k{}^2} + \frac{1}{2}\Big(1 + \frac{1}{\mu}\Big)\frac{1}{r_k} - \frac{1}{2\mu}\frac{1}{r_k{}^2} \tag{43}$$

Приравняем производную по **r** к нулю и найдем значение, соответствующее минимальному значению целевой функции при фиксированном наборе весов:

$$r_k = \frac{\mu {w_k}^2 + 1}{\mu + 1} \quad (44)$$

Алгоритм минимизации включает следующие шаги:

1. Выбор начальных значений для **r_in, w_in, c_in.**
2. Подбираем новые значения **w** по методу Ньютона, пока не выполнится условие: $\|grad(\boldsymbol{w\_new})\|$<eps
3. Вычисляем значение критерия **J**
4. Находим **r_new** по формуле (45)
5. Вычисляем значение критерия **J_new** при посчитанном **r_new**
6. Если не выполнилось условие |**J_new-J**|<eps, то считаем **c_new** (параметр меняется на фиксированную величину) и возвращаемся к пункту 2.

4.3.2 Подбор значений параметра регулируемой селективности

Множество параметров регулируемой селективости μ приводит к набору моделей, среди которых надо выбрать имеющую наилучшей обобщающей способностью.

Обычно параметры модели подбираются с помощью процедуры кросс-валидации. В нашей работе будем использовать кросс-валидацию по 10 блокам.

Обычно подбор параметров является достаточно трудозатратным процессом, т.к требуется строить модели «с нуля» для очень широкого диапазона значений μ.

Есть способ сокращения времени вычислений. Основная идея заключается в последовательном изменении параметра регуляризации и использовании коэффициентов модели, найденных при предыдущем значении параметров в качестве начального приближения новой модели. С точки зрения реализации, это, по сути, означает добавление внешнего цикла, где увеличивается параметр μ.

5 Экспериментальные исследования

5.1 Сравнение логистической регрессии с регулируемой селективностью с другими методами отбора признаков

Сравним модель логистической регрессии с регулируемой селективностью со стандартными методами отбора признаков (Ringe, Lasso, Elastic Net).

Для экспериментов была взята база German Credit Data. Заемщики описываются 20 атрибутами. В выборке 1000 записей, 700 положительных примеров и 300 отрицательных.

Перечислим признаки:

1. Состояние существующего счета (категориальный)
2. Срок кредита в месяцах (непрерывный)
3. Кредитная история (категориальный)
4. Цель получения кредита(категориальный)
5. Сумма кредита (непрерывный)
6. Состояние сберегательного счета (категориальный)
7. Стаж на текущем месте работы (категориальный)
8. Какой процент от дохода составляет выплата по кредиту (категориальный)
9. Семейное положение и пол (категориальный)
10. Наличие поручителей (категориальный)
11. Живет на данном месте в течение (категориальный)
12. Имущество (категориальный)
13. Возраст (непрерывный)
14. Наличие других невыплаченных кредитов (категориальный)
15. Тип жилья (категориальный)
16. Количество предыдущих кредитов в данном банке, включая открытые (категориальный)
17. Работа (категориальный)
18. Количество иждивенцев (категориальный)
19. Наличие телефона (категориальный)
20. Иностранный рабочий (категориальный)

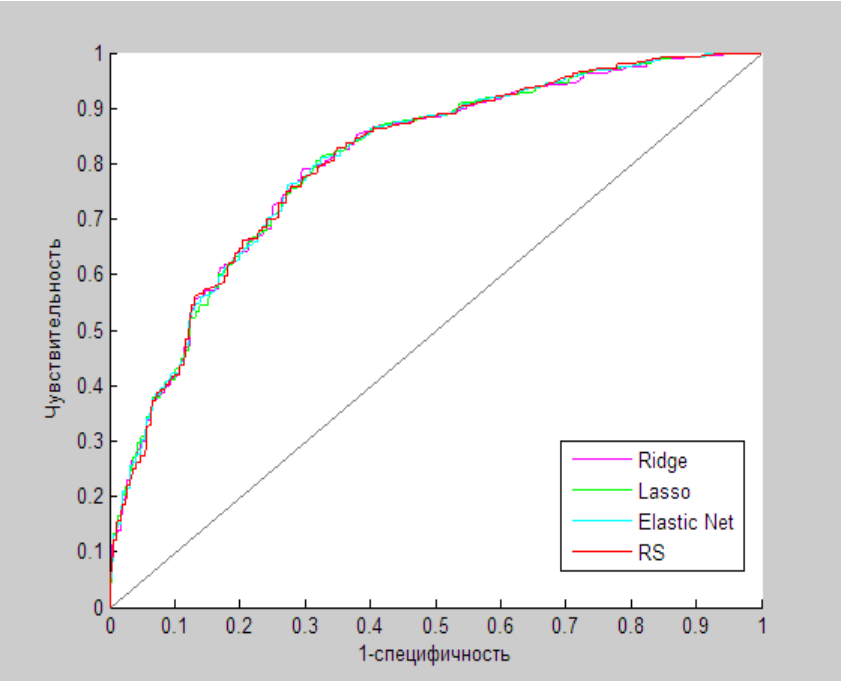

Рис 4. ROC-кривые, построенные по всем данным. Значения площади под ROC-кривой:AUG Ridge = 0.8206, AUG Lasso = 0.8177, AUG Elastic Net = 0.8169, AUG. Наш метод = 0.8188.

На полных выборках методы ведут себя практически одинаково. Теперь сократим размер обучающей выборки до 200 и 100 объектов и повторим эксперимент.

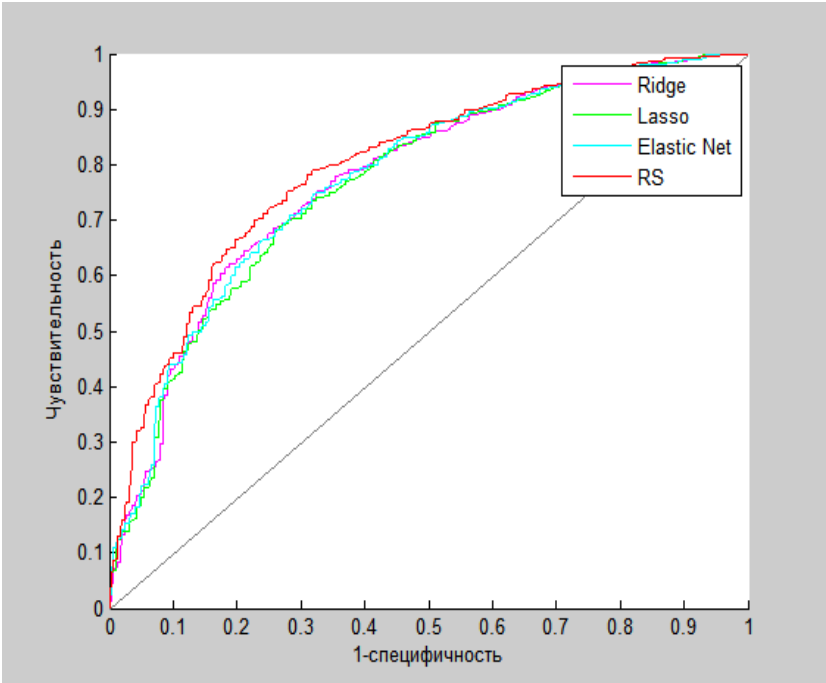

Рис 5. ROC-кривые при M=200 (German). Значение площади под ROC-кривой:AUG Ridge = 0.7749, AUG Lasso = 0.7689, AUG Elastic Net = 0.7752, AUG Наш метод = 0.7984.

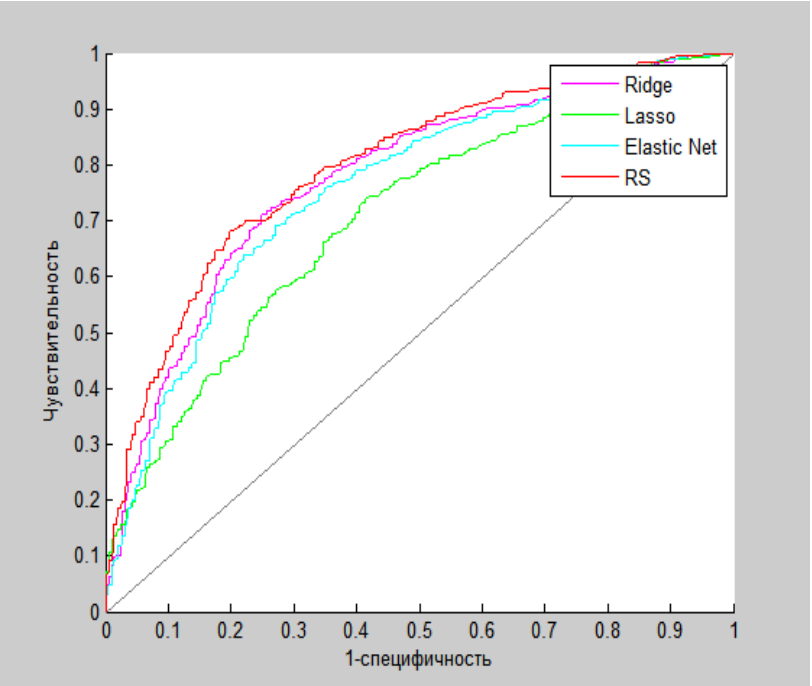

Рисунок 6. ROC-кривые при M=100 (German). Значение площади под ROC-кривой:AUG Ridge = 0.7765, AUG Lasso = 0.7092, AUG Elastic Net = 0.7593, AUG. Наш метод = 0.7963.

Получается достаточно интересный результат. Модель логистической регрессии с регулируемой селективностью лучше других методов работает на маленьких выборках.

5.2 Модель с экспертными ограничениями

Если выборка корректна, то разницы в качестве модели с и без ограничений практически нет. Но если берется некорректная, несбалансированная обучающая выборка, то наличие ограничений улучшает модель.

Для эксперимента брались данные ОТП-Банка, предоставленные для ММРО 2015. Имеется выборка из 15233 заемщиков, которые описываются 52 признаками.

Перечислим признаки, описывающие клиента:

1. уникальный идентификатор объекта в выборке
2. целевая переменная:отклик на маркетинговую кампанию (1 - отклик был зарегистрирован, 0 - отклика не было)
3. возраст клиента
4. социальный статус клиента относительно работы (1 - работает, 0 - не работает)
5. социальный статус клиента относительно пенсии (1 - пенсионер, 0 - не пенсионер)
6. пол клиента

7. количество детей клиента
8. количество иждивенцев клиента
9. образование
10. семейное положение
11. отрасль работы клиента
12. должность
13. форма собственности компании
14. отношение к иностранному капиталу
15. направление деятельности в нутри компании
16. семейный доход (несколько категорий)
17. личный доход клиента (в рублях)
18. область регистрации клиента
19. область фактического пребывания клиента
20. почтовый адрес область
21. область торговой точки, где клиент брал последний кредит
22. регион РФ
23. адрес регистрации и адрес фактического пребывания клиента совпадают(1 - совпадает, 0 - не совпадает)
24. адрес фактического пребывания клиента и его почтовый адрес совпадают(1 - совпадает, 0 - не совпадает)
25. адрес регистрации клиента и его почтовый адрес совпадают(1 - совпадает, 0 - не совпадает)
26. почтовый, фактический и адрес регистрации совпадают (1 - совпадают, 0 - не совпадают)
27. область регистрации, фактического пребывания, почтового адреса и область расположения торговой точки, где клиент брал кредит совпадают (1 - совпадают, 0 - не совпадают)
28. наличие в собственности квартиры (1 - есть, 0 - нет)
29. кол-во автомобилей в собственности
30. наличие в собственности автомобиля российского производства (1 - есть, 0 - нет)
31. наличие в собственности загородного дома (1 - есть, 0 - нет)
32. наличие в собственности котеджа (1 - есть, 0 - нет)
33. наличие в собственности гаража (1 - есть, 0 - нет)
34. наличие в собственности земельного участка (1 - есть, 0 - нет)
35. сумма последнего кредита клиента (в рублях)

36. срок кредита
37. первоначальный взнос (в рублях)
38. в анкете клиент указал водительское удостоверение (1 - указал, 0 - не указал)
39. в анкете клиент указал ГПФ (1 - указал, 0 - не указал)
40. количество месяцев проживания по месту фактического пребывания
41. время работы на текущем месте (в месяцах)
42. наличие в заявке телефона по фактическому месту пребывания
43. наличие в заявке телефона по месту регистрации
44. наличие в заявке рабочего телефона
45. количество ссуд клиента
46. количество погашенных ссуд клиента
47. количество платежей, которые сделал клиент
48. количество просрочек, допущенных клиентом
49. номер максимальной просрочки, допущенной клиентом
50. средняя сумма просрочки (в рублях)
51. максимальная сумма просрочки (в рублях)
52. количество уже утилизированных карт (если пусто - 0)

В выборке имеется много непрерывных признаков, которые в изначальном формате нельзя подавать в модель. Они были преобразованы в категориальные переменные и значения категорий были заменены на соответствующие WOE значения. Так же был проведен корреляционны анализ и анализ значений признаков (выбросы, экстремальные значения, анализ соотношений конечных классов признака относительно целевой переменной) перед построением модели. В результате этого признаковое пространство сократилось.

Эксперимент строился следующим образом:

1. строилась модель без ограничений
2. анализировались знаки получившихся коэффициентов
3. на некоторые коэффициенты, которые имели нелогичный знак, накладывались ограничения и модель перестраивалась

Проводились эксперименты со следующими версиями ограничений:

№ эксперимента	Положительные коэффициенты	Отрицательные коэффициенты
1	Должность Личный доход клиента (в рублях) Область фактического пребывания клиента	Средняя сумма просрочки (в рублях) Максимальная сумма просрочки (в рублях)
2	Должность Первоначальный взнос (в рублях) Время работы на текущем месте (в месяцах) Средняя сумма просрочки (в рублях) Максимальная сумма просрочки (в рублях)	Средняя сумма просрочки (в рублях) Максимальная сумма просрочки (в рублях)
3	Пол клиента Возраст клиента Наличие в собственности квартиры (1 - есть, 0 - нет)	
4	Наличие в собственности загородного дома (1 - есть, 0 - нет) Наличие в собственности гаража (1 - есть, 0 - нет) Наличие в собственности гаража (1 - есть, 0 - нет) Наличие в собственности квартиры (1 - есть, 0 - нет)	
5	Количество ссуд клиента Количество месяцев проживания по месту фактического пребывания Время работы на текущем месте (в месяцах)	

Таблица 3. Описание экспериментов

Результаты экспериментов:

№ эксперимента	AUC модели с ограничениями	AUC модели без ограничений
1	0.69584	0.69583
2	0.69583	0.69583
3	0.6959	0.69583
4	0.7029	0.69583
5	0.7051	0.69583

Таблица 4. Результаты экспериментов

Из Таблицы 4 видно, что модель с экспертными ограничениями работает либо лучше модели без ограничений, либо так же хорошо. Но при их использовании модель становится намного более интерпретируемой и логичной, что очень важно для задачи кредитного скоринга.

6 Заключение

В результате данной работы была разработана модель логистической регрессии с регулируемой селективности при наличии экспертных ограничений. Были представлены алгоритм обучения такого алгоритма и результаты экспериментов, а так же рассмотрены основные свойства модели. Модель хорошо подходит для банков, в которых мало данных, либо они некорректны. Так же, такая модель может быть применима для МФО, где всегда данные более искаженные, чем в банках.

Список используемой литературы

1. Вишняков, И.В. Методы и модели оценки кредитоспособности заемщиков / И.В. Вишняков - СПб.: СПбГИЭА, 1998. – 267 с.
2. Воронцов, К.В. Математические методы обучения по прецедентам (курс лекций) [Электронный ресурс] / К.В.Воронцов. – Режим доступа: http://www.machinelearning.ru/wiki/index.php
3. Кочедыков, Д. А. Система кредитного скоринга на основе логических алгоритмов классификации / Д.А. Кочедыков, А. А. Ивахненко, К. В. Воронцов // Математические методы распознавания образов-12.— М.: МАКС Пресс., 2005. – С. 349-353.
4. Abdou, H. A. Credit scoring, statistical techniques and evaluation criteria: A review of the literature / H.A. Abdou, J. Pointon J // Intelligent Systems in Accounting, Finance and Management. - 2011. – Т. 18. – 59-88 с.
5. Fan, J. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties / J. Fan, R. Li. // Journal of the American Statistical Association – 1996. – 1348–1360 с.
6. Glmnet for Matlab [Электронный ресурс] / J. Qian, T. Hastie, J. Friedman., R. Tibshirani, N. Simon. - 2013. – Режим доступа: http://web.stanford.edu/~hastie/glmnet_matlab/
7. Guyon, I. An introduction to variable and feature selection / I.Guyon, A. Elisseeff // The Journal of Machine Learning Research. – 2003. – Т. 3. – 1157-1182 с.
8. Zou H. Regularization and variable selection via the elastic net / H. Zou, T. Hastie // Journal of the Royal Statistical Society: Series B (Statistical Methodology). - 2005. - 301–320 с.
9. Hand, D. J. Statistical classification methods in consumer credit scoring: a review/ D. J. Hand, W. E. Henley //Journal of the Royal Statistical Society: Series A (Statistics in Society). – 1997. – Т. 160. – №. 3. – 523-541 с.
10. Hosmer, D. W. Applied logistic regression / D.W.Hosmer, S. Lemeshow . – John Wiley & Sons, 2004.
11. Friedman, J. Regularization Paths for Generalized Linear Models via Coordinate Descent / J.Friedman, T. Hastie, R. Tibshirani // Journal of Statistical Software. – 2010. - 1-22 с.
12. Liu, Y. Data mining feature selection for credit scoring models / Y.Liu, M. Schumann //Journal of the Operational Research Society. – 2005. – Т. 56. – №. 9. – 1099-1108 с.

13. Tibshirani, R. Regression shrinkage and selection via the lasso / R.Tibshirani // Journal of the Royal Statistical Society. Series B (Methodological). – 1996. - 267–288 с.

14. Somol, P. Filter- versus wrapper- based feature selection for credit scoring / P.Somol // International Journal of Intelligent Systems. – 2005. – Т. 20. – №. 10. – 985-999 с.

15. Statlog (Australian Credit Data) Data Set [Электронный ресурс]. - Режим доступа: https://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval

16. Statlog (German Credit Data) Data Set [Электронный ресурс]. Режим доступа: http://www.statistik.lmu.de/service/datenarchiv/kredit/kredit_e.html

17. Tang, J. Feature selection for classification: A review / J. Tang, S. Alelyani , H. Liu //Data Classification: Algorithms and Applications. Editor: Charu Aggarwal, CRC Press In Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. – 2014.

18. Thomas, L. C. Consumer Credit Models: Pricing, Profit and Portfolios: Pricing, Profit and Portfolios / L.C. Thomas. – Oxford University Press, 2009.

19. Tsai, C. F. Feature selection in bankruptcy prediction / C.F. Tsai //Knowledge-Based Systems. – 2009. – Т. 22. – №. 2. – 120-127 с.

20. Fan, J and Li R. (1996). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, pp. 1348–1360 с.