НОВ БЪЛГАРСКИ УНИВЕРСИТЕТ

Департамент по информатика

Автореферат

Компютърно симулиране и оценка на фирмени рискове

Автор:
Слав Емилов Ангелов

Научен ръководител:
проф. дмн Евгения Стоименова
Научен консултант:
доц. д-р Иван Костов

на дисертация за присъждане на образователна и научна степен "Доктор"
в област на висше образование 4. Природни науки, математика и информатика,
професионално направление 4.6. Информатика и компютърни науки

София
19 март 2019 г.

ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Обемът на дисертационния труд е 194 страници, в които са включени 41 фигури, 8 таблици и 1 приложение. Списъкът на използваната литература съдържа 144 източника – 140 на английски и 4 на български език.

Дисертацията има следната структура:

- Страница с цитат;
- Кратко представяне на английски;
- Съдържание;
- Списък на фигурите;
- Списък на таблиците;
- Съкращения;
- Уводна глава - 38 страници;
- Глава 2: „Подбор на данни и методи за изследване“ - 41 страници;
- Глава 3: „Моделиране на финансовото представяне на газоразпределителния сектор в България“ - 33 страници;
- Глава 4: „Регресионна техника за редуциране на влиянието на зависимите предиктори“ - 16 страници;
- Глава 5: „Крос-валидирана постъпково конструирана множествена регресия“ - 17 страници;
- Глава 6: „Използване на резултатите при оценка на риска“ - 20 страници;
- Глава 7: „Заключение“ - 5 страници;
- Приложение - 1 страница;
- Библиография.

Дисертантът е редовен докторант към департамент „Информатика“ на Нов български университет – гр. София.

Съдържание

Съдържание ii

1 Увод 1
- 1.1 Цели и задачи . . . 4
- 1.2 Актуалност . . . 5
- 1.3 Приложимост . . . 6
- 1.4 Обобщение на Глава 1 . . . 7

2 Подбор на данни и методи за изследване 8
- 2.1 Полуавтоматизирана процедура за конструиране на регресионен модел . . . 8
- 2.2 Обобщение . . . 10

3 Моделиране на финансовото представяне на газоразпределителния сектор в България 11
- 3.1 Газоразпределителният сектор в Република България . . . 11
- 3.2 База за прогнозиране на финансовото представяне . . . 12
- 3.3 Построени модели . . . 12
  - 3.3.1 ROA модела . . . 13
  - 3.3.2 FL модела . . . 15
- 3.4 Обобщение . . . 16

4 Регресионна техника за редуциране на влиянието на зависимите предиктори 18
- 4.1 Техника за справяне със зависими предиктори . . . 19
  - 4.1.1 Теоретично описание на базата за предложената техника . . . 19
  - 4.1.2 Създаване на компоненти . . . 20
- 4.2 Алгоритъм на техниката в случая на евристичен подход чрез t-стойности . . . 21
- 4.3 Обобщение . . . 22

5 Крос-валидирана постъпково конструирана множествена регресия 23
- 5.1 Същност на метода . . . 23
- 5.2 Алгоритмично представяне на „ядрото“ на техниката . . . 25
- 5.3 Методът при повече от два предиктора . . . 26
- 5.4 Обобщение . . . 27

6 Използване на резултатите при оценка на риска 29
- 6.1 Оценка на риска при фирмите от газоразпределителния сектор . . . 29
- 6.2 Обобщение . . . 32

7 Заключение 33
- 7.1 Научни и приложни приноси . . . 34
- 7.2 Апробация на резултатите . . . 36

Библиография 38

Глава 1

Увод

Кризата на глобалната икономика от 2008 година причини масови поражения на световно ниво. Актуалността на компютърното симулиране и оценяване на фирмени рискове произтича от тези поражения, фалити и изключвания на фирми, от борсовите загуби на пенсионни и осигурителни фондове, на застрахователни дружества и спасяване на институции. Случилите се поражения до голяма степен се дължат на непознаване на рисковете и липсата на достатъчна готовност от страна на фирмите и заинтересованите лица да реагират на рискове.

В отговор на щетите от финансовата криза, след година усилия на множество специалисти в областта, бе разработен международен стандарт за управление на рискове ISO 31000:2009. Той съдържа в себе си дефиниции на понятията, с които трябва да се борави в сферата на управление на рисковете и обобщава на високо ниво всичко, което дисциплината управление на рисковете трябва да съдържа. Високото ниво на професионализъм при изготвянето му и всеобщата готовност за приемане е причината този стандарт да залегне като начална точка в този дисертационен труд. Стандартът определя следната дефиниция за риск:

*„Организациите, независимо от техния вид и големина, са изправени пред вътрешни и външни фактори и влияния, които създават неувереност дали и кога ще постигнат своите цели. Влиянието на тази неопределеност по отношение на постигането на целите на дадена организация представлява т. нар. „**риск**”.”*

Подходите към фирмените рискове, които макар уникални като моделиране, измерване и оценка за всяка фирма, предполагат да бъдат развити редица общи методологии за управление на риска. Според ISO 31000 управлението на риска следва да спазва определени точно дефинирани принципи и рамка. Неразделна част от дефинираната рамка е процеса на управление на рисковете. Това се явява работа по точка 5 от ISO 31 000. Ядрото на процеса на управление на рисковете се състои от следните

три стъпки - установяване на обстоятелствата; оценяване на риска (идентифициране на риска, анализ на риска, придаване на стойност на риска); въздействие върху риска. Обекта на изследване на този труд засяга тези три фундаментални стъпки с акцент върху установяване на обстоятелствата и оценяване на риска.

В практиката съществува голямо множество различни фирмени рискове. Всяка една дейност може да бъде разглеждана като риск. Анализирането на всеки риск е трудоемък процес и като начало изисква по-строга систематизация на това какво е риск и какви рискове съществуват. Една от създадените систематизации за видовете рискове и техните източници, се представя чрез класификацията на Фигура 1.1. Тази класификация е базирана на FIRM Risk Scorecard системата, разработена е от водещите организации по управление на риска във Великобритания (the Association of Insurance and Risk Managers (AIRMIC), the public sector risk management association (Alarm) и the Institute of Risk Management (IRM)) и е представена на английски в един от производните от стандарт ISO 31 000 документи - [AIRMIC, 2010]. Източниците на риск са многобройни и качествено различни. Всеки от упоменатите външни и вътрешни източници на Фигура 1.1 може да се разпадне на подизточници, допълнително самата класификация не претендира да е изчерпателна и е възможно в конкретни случаи да има други източници и рискове. Това прави процеса на идентифициране и оценяване на рисковете трудоемък и субективен. Изработването на помощни инструменти и техники, които да идентифицират качествено и количествено рискове и източници на високо ниво на обобщение, би могло да намали субективността при идентифицирането и оценка на рисковете, а също и да облекчи работата върху целия процес. Помощни инструменти и техники, базирани на научните сфери статистика и машинно обучение, са приоритетни цели на това дисертационно изследване. Допълнително конкретизиране на целите ще бъде направено в Раздел (1.1), но преди това следва да завършим постановката.

От Фигура 1.1 може да се придобие представа какво голямо множество от външни и вътрешни източници на риск съществува за всяка една икономическа единица дори и на високо ниво на обобщение. Ако искаме допълнително да опростим схемата, може да подходим от следната гледна точка. Всеки един от източниците на риск при проявяването си ще повлияе пряко или косвено на финансовото състояние на дадена компания (икономическа единица). Това е факт независимо дали възникналото явление е можело да бъде количествено оценено или не и независимо от това дали изобщо е отчетено съществуването и настъпването на това явление. Всяко едно събитие, което настъпва в ежедневието на една компания, се приспада към някой от източниците на риск и участва пряко или косвено във финансовия баланс на дадената компания. Това дали в дългосрочен план финансовия баланс е положителен определя дали фирмата

Фигура 1.1: Класификация на рисковете и техните източници на база на FIRM Risk Scorecard системата.

ще остане на пазара или не. Така на най-високо ниво остава един единствен риск - риска от фалит. Съществува голямо множество от статии моделиращи риска от фалит, преглед върху които ще бъде направен в литературния обзор към дисертацията. За моделирането на риска от фалит се използват широко множество от статистически подходи и техники за машинно обучение. Общото между тях е, че използват финансови коефициенти, предвиждат дали фирмата ще фалира и кога. Произведените модели са универсални по отношение на икономическата ниша, но обикновено изключват банковия сектор и комуналните услуги. От логически съображения разглеждането на конкретен сектор по отношение на риска от фалит би дало по-добри резултати от прилагането върху него на модел, предназначен за произволна фирма. Проблем при разработването на модел, прогнозиращ риска от фалит върху конкретен сектор е, че е много вероятно да не се разполага с достатъчно данни на фалирали спрямо функциониращите фирми. Това поражда въпроса дали не може да се премине към друг подход относно риска за фалит, който да извлича информация полезна при процеса на управление на рисковете. Моделите за оценка на риска от фалит обикновено се

базират на няколко високоинформативни финансови коефициенти. Моделирането на избрани високоинформативни финансови съотношения би разкрило от какво зависят самите те. Така се оформя верига, в която риска от фалит зависи предимно от няколко коефициента. Тези няколко коефициента зависят от други финансови съотношения и някои величини. В тази верига, ако не може да бъде оценен директно риска от фалит, то може да бъдат установени, по финансови съотношения в конкретния модел, други подфактори, които му влияят. Така се оформя една свързана система, в която тръгвайки от финансовите съотношения, моделиращи риска от фалит, може да се достигне до допълнителни детайли относно поведението на конкретна група от фирми и събитията, от които тези фирми най-силно се влияят. Колко резултатен може да бъде подход от този тип, ще бъде демонстрирано върху конкретно избран сектор - газоразпределителните фирми в РБ. Допълнително към поставената цел в този дисертационен труд се включва разработката на похвати, методи и алгоритми, които да служат за спомагателни инструменти при изготвяне на модели.

1.1 Цели и задачи

Един от най-много разискваните и описвани рискове, в литературата и практиката, е рискът от фалит. При моделирането на това явление, на база данни и съображения, се установява дали няма да настъпи, дали ще настъпи и ако настъпи, то кога. Отговорите на тези въпроси, дори и в случаите на високо ниво на достоверност, дават сравнително малко информация и в съвременните пазарни условия редица потребители биха били заинтересовани от по-високо ниво на детайли. Фалитът се явява следствие на множество успоредно вървящи и взаимно свързани процеси. Такива процеси са постепенното намаляване на печалбите и увеличаване на задлъжнялостта на дадена фирма. Предвиждането на всеки един гореупоменат процес поотделно се явява задълбочаване нивото на информация, което в редица случаи води до по-добро справяне с неопределеността.

В този труд се разглежда статистическо моделиране и интерпретиране на показатели на фирми, заемащи обща икономическа ниша. **Основната цел е** създаването на спомагателен инструмент, който да се използва за третиране на неопределеността по отношение на риска от фалит или финансовото здраве на компания от конкретен икономически сектор. Създаденото следва да е подходящо за база на система за подпомагането на решения (decision support system) по отношение на оценка на риска в газоразпределителния сектор, а евентуално негова модификация да е подходяща за

произволен икономически сектор или за друг конкретен. За постигането на тази цел обозначаваме следните задачи:

- Изготвяне на съображения и принципи за избор на ключови по отношение на риска от фалит ФС за моделиране;
- Подбор на методи и техники за обработка и анализ на данни;
- Създаване на полуватоматизиран подход за моделиране на избраните ключови ФС на база минали наблюдения (данни за финансови съотношения, променливи специфични за даден отрасъл, макроикономически величини) с възможност за актуализация;
- Създаване на подход за анализ на конструираните модели и използването им за извличане на изводи за целия отрасъл;
- Създаване на ориентировъчни критерии и похвати за определяне на финансовото здраве на компания;
- Създаване на техники за машинно обучение (machine learning) за подпомагане на процеса на статистическо моделиране с акцент върху прогнозите.

Конкретен обект за анализ в този труд са компаниите в РБ, които притежават лиценз за продажба и разпределение на природен газ, няколко дузини частни фирми и две държавни. Газоразпределителният сектор е специфичен и универсалните модели за оценка на риска от фалит подвеждат в своите изходни стойности. Допълнително универсални методи са неуместни и поради спецификите на българската икономика. Тези факти обуславят необходимостта от конкретно изследване за сектора. Информацията за целите на изследването взимаме от задължителните годишни финансови отчети. Допълнително ще разработим и включим в анализа променливи специфични за конкретният отрасъл, отчитаме и макроикономически показатели, оказващи влияние върху стопанската среда.

Акомпаниращи средства за постигане на целите ще са езика за статистическа обработка на данните R и MS Excel.

1.2 Актуалност

Поставените цели са от универсален характер и са част от икономическата действителност на 21 век. Създаването на инструменти, които спомагат за изследването на това дали една фирма (от даден отрасъл) ще расте или ще изпадне в неплатежоспособност е в интерес на: акционери, мениджъри, икономисти, банки, политици, държавни лидери, институции, застрахователи, настоящи и бъдещи служители. За всички тях е важно неопределеността относно риска от фалит на конкретна икономическа единица

да бъде сведена до минимум. Допълнително при анализ на риска от фалит се изследват процесите, които протичат в съответните фирми и сектора като цяло. Това води до едно по-добро познание върху средата и процесите, което може да се използва за увеличаване на ефективността и ефикасността на фирмите.

Конкретно фирмите от газовия сектор, както и всеки друг икономически отрасъл, са обект на интерес по отношение на гореупоменатото. Още повече, газовият сектор се явява част от енергийния сектор, което засяга пряко националната сигурност. Това поставя проблематиката на ново ниво и дава приоритет на този отрасъл като базов. Вникване в икономическата действителност на фирмите от този сектор допринася към по-качествените държавни регулации в това направление, а може и да спомогне за по-оптимален мениджмънт на отделните компании.

1.3 Приложимост

Всеки един ползвател на информационна услуга е заинтересован тя да бъде своевременна, удобна, точна и ясна. Конкретно за фирмите от газовия сектор моделираните изходни показатели са два, описващи съответно фирмената производителност и задлъжнялост. Те могат да бъдат получени за конкретна фирма, за произволна група от фирми или за целия сектор. Тези два показателя са достатъчни, за да се придобие представа дали дадена икономическа единица ще се развива в следващия времеви период или върви назад. Актуализацията на моделите за бъдещи цели става чрез въвеждане на нови наблюдения за фирмите и повторно оценяване на моделите. В разработката е взето предвид съображението, че актуализацията на моделите трябва да протича по възможно най-прост начин. На база на моделираните показатели са изведени ориентировъчни критерии и похвати за оценка на риска от фалит. Цялостната разработка е концептуална и може лесно да се модифицира или използва съвместно с допълнителни инструменти и експертни съображения. Всичко е описано постъпково и детайлно с цел разработката да е имплементируема в система за подпомагане взимането на решения по отношение на риска от фалит.

Технологичната рамка, представена в дисертацията е приложима за произволен отрасъл и фирма от него. Самата конструкция на изложените методи позволява дообогатяване и развиване. Данните необходими за започване на дадено изследване, както вече се спомена, са достъпни от съответните държавни институции. Необходимата информация е съхранявана и обработвана в Excel и в последствие конвентирана в CSV входен файл за езика R , който е свободен за ползване и световно утвърден по отношение на вероятности и статистика. Самите използвани статистически техники и

похвати, макар и комбинирани в по-сложна полуавтоматизирана процедура, изискват стандартни математически познания за разбиране и употреба, позволявайки разпределяне на по-голямата част от работата сред хора широки специалисти, което допринася за производителността на разработването на нови изследвания върху избран икономически сектор.

В процеса на работа изработихме две техники за машинно обучение за намаляване на грешката от множествена регресия при определени условия. Тези техники са универсални и могат да се прилагат във всички сфери, в които се прилага и многомерния регресионен анализ.

1.4 Обобщение на Глава 1

В уводната глава излагаме серия от въвеждащи раздели. Дефинирахме задачата - моделиране на ключови ФС за оценка на риска от фалит в българския газоразпределителен сектор. Показахме приложността и актуалността на поставените цели - изследването може да се използва като спомагателен инструмент за оценка на риска от фалит от широко множество заинтересовани лица. Описахме и постигнатите от нас приноси - добавена стойност в оценка на риска от фалит в газоразпределителния сектор на България и два нови алгоритъма за оценка на множествена регресия. Въведохме читателите в сферата на финансовите съотношения и моделите за оценка на риска от фалит. Избрахме две финансови съотношения измерители на печалбата и задлъжнялостта като ключови за дисертационния труд относно оценка на рисковете. В последствие разгледахме широко множество от статистически техники за прогнозиране и избрахме множествена регресия за основен метод за постигане на целите ни. Завършихме главата с преглед на съществуващи разработки в сферата на газоразпределителния сектор на България и заключихме, че изследванията са сравнително оскъдни, а по конкретно поставените цели такива липсват.

Глава 2

Подбор на данни и методи за изследване

В тази глава излагаме голяма част от теоретичното описание за процеса на събиране, обработка, анализ и обобщаване на необходимата информация за постигане на конкретните цели на този дисертационен труд. Обясняват се накратко средствата на счетоводната отчетност и избраните статистически похвати. Излагаме и полуавтоматизирана система за конструиране на множествена регресия. Практическата реализация на описаното се намира в Глави 3 и 6, но косвено се използва и в останалите глави.

2.1 Полуавтоматизирана процедура за конструиране на регресионен модел

В този раздел ще обобщим всички съображения по отношение на множествената регресия в полуавтоматизирана процедура за конструиране на модели. Предлагаме следната процедура:

1. Обработка на данните:

   (а) Зареждаме данните - проверка за брой редове, стълбове, имена на променливи, типове данни;
   (б) Проверка за липсващи стойности - изтриват се редовете с липсващи стойности или се попълват липсите (не се препоръчва);
   (в) Онагледяваме два по два предикторите спрямо моделираната променлива посредством скатърплот. Целта е евентуално набелязване на предиктори, които имат нелинейна зависимост с моделираната променлива. Възможно е и да се локализират грешки в данните;

(г) НЕ се изтриват изключителни наблюдения освен, ако те не са установени грешки в данните.

2. Избор на предиктори:

(а) Пускаме множествена регресия между моделираната променлива срещу всички предиктори - ако броят на предикторите е по-голям от броя на наблюденията, то първо отсяваме някои от предикторите, които имат висок коефициент на определеност с останалите предиктори (например над 0.9 поне с един предиктор), в последствие пускаме множествената регресия;

(б) Проверяваме тази регресия за влиятелни наблюдения и ги третираме - например с разстояния на Кук, ако са над 1, то изтриваме наблюдението (може и да го манипулираме, например чрез претегляне);

(в) Проверка за неконстантна вариация на грешката - скатърплот остатъци срещу предвидени стойности, няма нужда от по-прецизни статистически тестове на този етап. При установяване на значителен проблем отиваме на Стъпка 4 за трансформация;

(г) С вече третираните данни за неконстантна вариация на грешката и влиятелни наблюдения прилагаме PLSR, за да се ориентираме колко добро R^2 да очакваме за модел за конкретен брой предиктори;

(д) Започваме избор на модели - пускаме последователно froward, backward, stepwise, allsubsets регресии с избрания лимит за броя на предикторите. Ако поради голям брой предиктори не може да пуснем allsubsets регресия, то първо отсяваме достатъчно предиктори с backward регресия и след това прилагаме allsubsets. Избраните по количествени (s^2, R^2, брой предиктори, разпределение на грешката, корелации между предикторите, p-стойности, други) и качествени съображения (смисъл на предикторите, съображения спрямо конкретната цел и други) модели преминават към Стъпка 3;

3. Диагностика на отсятите модели:

(а) Проверка за влиятелни точки - наблюдения с разстояния на Кук над 1 се премахват или се претеглят;

(б) Проверка за неконстантна вариация на грешката - скатърплот остатъци срещу предвидени стойности, скатърплотове за локализиране на влиянието на отделните предиктори върху вариацията на грешката, теста на Брюш-Паган. При наличие на неконстантна вариация се третира проблемния(ите) предиктор(и) чрез трансформация с подходяща функция;

(в) Проверка за нелинейност с *Компонент плюс остатък плотове*. Ако се установи нелинейност по даден предиктор, то той отново може да се подложи на трансформация с функция;

(г) Проверка за мултиколинеарити - VIFs над 4 и корелации между предикторите над 0.7 може да са проблемни, а VIFs над 10 и корелации над 0.9 следва да се анализират и третират (в краен случай чрез отстраняване на проблемен предиктор). Ако модела е за прогнозни цели това дали мултиколинеарити проблема е сериозен или не може да се валидира чрез тестови множества и измерване на средноквадратичната коренувана грешка.

(д) При необходимост премахваме и свободния коефициент - по изключение той може да се запази дори и при p-стойност над 0.2, но не и при p-стойности над 0.5;

(е) Тестване на представянето на модела - каква е средната грешка на модела за бъдещи наблюдения се тества чрез тестово и обучаващо множество или чрез крос-валидация, ако променливите са малко;

4. Трансформации - Стъпки 2 и 3 могат да се повторят след прилагането на една от следните трансформации:

(а) Логаритмуване само на моделираната променлива;
(б) Логаритмуване само на предикторите;
(в) Логаритмуване на всички променливи;
(г) Прилагане на специфична функция само върху конкретна променлива или променливи - използваната функционална форма се избира на база на скатърплота между съответния предиктор и моделираната променлива;
(д) Добавяне на допълнителни предиктори към множеството данни, които представляват степени на съществуващи предиктори от множеството (полиномна регресия) - степени над 3 са силно Не препоръчителни.

Процедурата е имплементируема на произволен език за програмиране, но на езика R са вече реализирани всички необходими функции. В приложението към дисертационния труд има препратки към файлове с реализация на Стъпки 2 и 3 от процедурата в R.

2.2 Обобщение

В тази глава излагаме основната технологична рамка, която е необходима за боравене с данните за постигане на целите на това дисертационно изследване. Процесът е проследен от ниво търсене и събиране на данните; през метода за моделиране и неговите особености; до инструменти за реализация на самите модели и изчисления; и интерпретиране на получените модели. Предложихме и процедура за конструиране на регресионни модели. Изложената методология е напълно приложима за множество задачи извън този дисертационен труд, но е конкретно насочена да служи като база за показаното в следващите глави.

Глава 3

Моделиране на финансовото представяне на газоразпределителния сектор в България

В тази глава ще се запознаем с газоразпределителния сектор в Република България, ще се опишат негови специфики от гледна точка на статистическото моделиране, ще се конструират и диагностицират модели, целящи да покажат финансовите процеси, които протичат в сектора тук и сега, и в близко бъдеще.

3.1 Газоразпределителният сектор в Република България

Разгледани са всичките 29 лицензирани фирми за 2014 година. В това число не са включени гигантите „Булгаргаз" ЕАД и „Булгартрансгаз" ЕАД, които са държавна собственост, но счетоводните им отчети са анализирани, защото съдържат обобщаващи за сектора данни. Допълнително множество от изводите, които ще направим за частния сектор ще са валидни и за държавните компании като компании от сектора. Частните фирми са обект на изследването, защото неопределеността при тях е по-голяма, обхващат голям и разнообразен дял от пазара, а мащабно изследване върху тях не е известно. При разглеждане на отчетите се установява, че този сектор все още е в интензивно развитие у нас. Извършват се сливания, масивни инвестиции за газопроводи и някои от фирмите взимат значителни инвестиционни заеми. Две от дружествата все още са неактивни, което понижава броя на разглежданите дружества на 27. Четири от тях са без обявена счетоводна политика и приложения към отчетите,

но две от тях са включени в анализа. Друго е фалирало и не е включено. Други две са слети и се разглеждат като едно.

3.2 База за прогнозиране на финансовото представяне

Финансовото представяне на фирмите от газоразпределителния сектор в България ще бъде моделирано посредством два високоинформативни финансови коефициента - ROA и FL. Възвръщаемостта на активите ROA е измерител на това колко печалба е способна да генерира една фирма от всичко, което притежава. Съотношението ROA дефинираме в следната форма:

$$ROA(t) = \frac{\text{Нетен доход}(t) + 0.90 * \text{Лихви}(t)}{\text{Средногодишни активи}(t)}, \tag{3.1}$$

където:

- t е конкретния период от време;
- Коефициентът 0.9 прилага *Данък общ доход*, който след 2006 година за България е 10%;
- Средногодишни активи$(t) = \frac{\text{Активи}(t)+\text{Активи}(t-1)}{2}$.

Много е важно да се отбележи, че Лихви не следва да се облагат с данък, ако фирмената печалба преди облагане с данъци плюс Лихви е отрицателно число. Ако фирмената печалба е отрицателна, но става положителна след добавяне на Лихви, то се облага с данък само положителния сбор.

Вторият високоинформативен ФС е FL или *Финансовия левъридж* на една фирма, който показва каква част от фирмата е финансирана от чужд капитал или с други думи колко е задлъжняла фирмата.

$$FL(t) = \frac{\text{Пасиви}(t) - \text{Кеш}(t)}{\text{Пасиви}(t) - \text{Кеш}(t) + \text{Собствен капитал}(t)} \tag{3.2}$$

3.3 Построени модели

Крайните модели, след процедурата за избор на предиктори и всички останали диагностики и валидации, са изложени в този раздел. Моделът, моделиращ ROA за удобство е наречен *ROA модел*, съответно в следващия раздел имаме и *FL модел*.

3.3.1 ROA модела

Финалният модел за ROA в година t е:

$$ROA(t) = \alpha_1 * PTA(t-1) + \alpha_2 * DC(t-1) + \alpha_3 * FL(t-1) + \alpha_4 * PTL(t-1) + \\ + \alpha_5 * LAOR(t-1) + \alpha_6 * I_{gas}(t-1) + \alpha_7 * Firm.size(t-1), \quad (3.3)$$

където:

- ROA е възвръщаемостта на активите, дефинирана във Формула 3.1;
- α_1,...,α_7 са регресионните коефициенти за оценяване;
- PTA = (Обезценки + Нетен доход + Амортизация) спрямо Средногодишни активи. Това е едно от ФС, показващи колко печалба може да генерира дадена фирма от активите си;
- DC = Амортизация спрямо Разходи. Това съотношение показва колко значителна е ролята на разходите за амортизация спрямо всички разходи;
- FL = (Пасиви - Кеш) спрямо (Пасиви - Кеш + Собствен капитал). Измерва каква част от фирмата е финансирана с дълг;
- PTL = (Обезценки + Нетен доход + Амортизация) спрямо Средногодишните пасиви. ФС от тип доход-дълг;
- $LAOR$ = (Средногодишни дълготрайни активи) спрямо Приходи от основна дейност. Това съотношение свързва физическата инфраструктура на дадена газоразпределителна фирма с приходите на съответната фирма;
- I_{gas} - индекс на цената на газта;
- $Firm.size$ измерва големината на фирмата по размера на активите и продадените количества газ.

Самото оценяване на модела е посредством функцията *lm()* в езика R, резултатите са показани на Фигура 3.1. ROA моделът е със сравнително високо R^2 и коефициентите му са много добре оценени (малки грешки в сравняване със съответните им оценки).

Изведеният модел е способен да прогнозира една година напред на база на входни данни от настоящата година. Допълнително, моделът онагледява основни фактори, влияещи на фирмения доход. На Фигура 3.1 могат да се видят статистики, характеризиращи модела. От тези наблюдения извеждаме някои емпирични заключения:

- При фирмите от газоразпределителния сектор на България се наблюдава икономия от мащаб;
- Покачването на цените на газта влияят негативно на печалбите в сектора;

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.041626 -0.009079 -0.000814  0.008360  0.044911

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
PTA         0.630438   0.055520  11.355  < 2e-16 ***
DC         -0.556306   0.101120  -5.501 2.55e-07 ***
FL          0.036597   0.006910   5.296 6.28e-07 ***
PTL        -0.031903   0.007692  -4.148 6.72e-05 ***
LAOR        0.013227   0.003257   4.061 9.27e-05 ***
GP.index   -0.019586   0.004277  -4.579 1.26e-05 ***
Firm.size   0.008315   0.002568   3.238   0.0016 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 0.0155 on 108 degrees of freedom
Multiple R-squared:  0.8753,    Adjusted R-squared:  0.8673
F-statistic: 108.3 on 7 and 108 DF,  p-value: < 2.2e-16
```

Фигура 3.1: Статистики за ROA модела, посредством lm() функцията в езика R.

- Финансовият левъридж увеличава печалбите;
- Съотношението DC влияе негативно на ROA. DC може да се каже, че е приблизително Амортизация спрямо Разходи. Но амортизационните ставки се контролират от избора на счетоводна политика на всяка една от фирмите и като следствие фирмите могат да влияят на измерителите на печалбата си;
- Съотношението $LAOR$ допринася положително към ROA. То показва, че фирма, която е инвестирала много във физическа инфраструктура в сравнение с приходите си се очаква да има по-големи печалби в бъдеще;
- PTA участва с положителен знак в ROA модела. $PTA(t-1)$ влияе най-значимо над $ROA(t)$ спрямо останалите предиктори. Тази релация показва, че печалбите в следващия времеви период зависят много от способността на фирмата да генерира печалби от своите Средногодишни активи сега.
- Най-трудна за интерпретация е фактора PTL. Това ФС показва всичко, което може да бъде отнесено към доход спрямо всичко, което може да бъде прието за дълг. То е с отрицателен знак в ROA модела. PTL показва, че размера на фирмените печалби спрямо размера на фирмения дълг е съотношение, което засяга бъдещите печалби. PTL в ROA модела показва, че ако процентната печалба на дадена фирма се увеличи през период t тогава и дълга трябва да се повиши, за да се очакват допълнителни процентни увеличения на печалбите в период $t+1$.

3.3.2 FL модела

Финалният модел за FL в година t е:

$$FL(t) = \beta_0 + \beta_1 * FA(t-1) + \beta_2 * BL(t-1) + \beta_3 * LATA(t-1) + \\ + \beta_4 * AL2(t-1) + \beta_5 * GRC2(t-1) + \beta_6 * LLE(t-1) + \beta_7 * EGP(t-1), \quad (3.4)$$

където:

- FL е финансовия левъридж или избраното ФС, измерител на фирмената задлъжнялост, вижте Формула 3.2;
- β_1,...,β_7 са регресионните коефициенти за оценяване;
- FA = Собствен капитал спрямо (Собствен капитал + Пасиви) или какъв е дела на собствения капитал в цялостното финансиране на фирмата;
- BL = (Текущи активи - Материални запаси) спрямо Краткосрочни задължения;
- $LATA$ = Дългосрочни активи спрямо Активи, показва колко е дела на дългосрочните активи от цялото;
- $AL2$ = Средногодишен кеш спрямо Средногодишни краткосрочни задължения;
- $GRC2$ = (Нетен доход + 0.9*Лихви) спрямо (Средногодишни дълготрайни активи + Средногодишни текущи активи - Средногодишни кр. задължения);
- LLE = Дългосрочни задължения спрямо Собствен капитал;
- EGP - съотношение цени газ-електричество.

Оценениият модел от (3.4) може да се види на Фигура 3.2. Моделът е с близък до максималния R^2 и добре оценени коефициенти по отношение на t-стойностите.

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.144637 -0.021245 -0.004122  0.020023  0.148127

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.769975   0.054623  14.096  < 2e-16 ***
FA          -1.119858   0.030695 -36.483  < 2e-16 ***
BL           0.030202   0.004830   6.253 9.04e-09 ***
LATA         0.278104   0.050527   5.504 2.69e-07 ***
AL2         -0.068866   0.013389  -5.144 1.28e-06 ***
GRC2        -0.385783   0.104078  -3.707 0.000339 ***
LLE         -0.031888   0.009432  -3.381 0.001018 **
EGP          0.023901   0.009183   2.603 0.010597 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 0.04063 on 104 degrees of freedom
Multiple R-squared:  0.9714,    Adjusted R-squared:  0.9695
F-statistic: 505.4 on 7 and 104 DF,  p-value: < 2.2e-16
```

Фигура 3.2: Статистики на FL модела, посредством lm() функцията в езика R.

Редица емпирични изводи могат да се направят от оценката на Уравнение 3.4:

- FA съотношението влияе отрицателно на задлъжнялостта. Това показва, че колкото повече една фирма е финансирана със собствен капитал в период t, толкова по-малко дълг се очаква тя да има в период $t+1$;
- Фирмите, които държат големи количества Кеш в сравнение с тяхните краткосрочни задължения се очаква да намалят задлъжнялостта си в следващия отчетен период, вижте $AL2$ съотношението;
- Фирмите, които се опитват да покрият краткосрочните си задължения с текущите активи се очаква да увеличат дълга си, което показва, че тази стратегия влошава финансовата им стабилност (вижте BL съотношението);
- $LATA$ съотношението показва, че за големи инфраструктурни проекти (разширяване на газоразпределителната мрежа и други) фирмите са склонни да взимат заеми;
- EGP съотношението индикира, че когато има благоприятна икономическа обстановка (цената на газта е значително по-малка от цената на електричеството) фирмите увеличават своя дълг. Разумно обяснение на този процес е, че фирмите се опитват да се разрастват по-бързо, когато има условия;
- LLE показва, че нарастването на дълга в период t повлиява негативно по-нататъчното разрастване на дълга в период $t+1$;
- $GRC2$ участва с отрицателен знак. От това може да бъде заключено, че ако фирма генерира все повече печалба от физическите си активи, то тогава ще започне да намалява своя финансов левъридж.

3.4 Обобщение

В тази глава представихме два регресионни модела, измерващи ключове финансови съотношения, измерители на представянето на фирмите от газоразпределителната индустрия в България. Един, измерващ дохода на фирмите (ROA модела) и един за дълга (FL модела). И двата модела съдържат макроикономически величини, което показва ползата от комбиниране на финансови съотношения с макроикономически величини, което не е често срещана практика. Двата произведени модела са с отлични характеристики - стабилни във времето, прецизно оценени и с добри резултати по отношение прогнозиране на бъдещи наблюдения. ROA моделът е способен да обясни 83% от вариацията в ROA съотношението, а FL моделът обяснява около 97% от вариацията в FL съотношението. Постигнатата висока обяснителна способност показва възможността да се моделира финансова и макроикономическа информация от сектора. Всички коефициенти от моделите са със статистическа значимост поне 98%, което е свидетелство

колко добре са оценени моделите. Моделите преминаха успешно и всички валидиращи тестове, което е индикатор за тяхната робастност. Грешките на двата модела при прогнозиране на нови наблюдения позволяват прогнозите им да са надеждна база за предупредителни сигнали по отношение на бъдещ банкрут или влошено състояние на дадена фирма. Същите тези ранни предупредителни знаци могат да се допълнят и с доверителни интервали (за предвидените стойности или наблюдаваните такива) за още по-надеждно определяне на риска от фалит. Посредством мин-макс трансформацията се позволи и степенуването по сила на влияние на предикторите като фактори, влияещи върху печалбата и задлъжнялостта на фирмите от сектора. На база на знаците пред коефициентите на предикторите се потвърждават някои основни допускания от сферата на счетоводството и финансите за фирми от газоразпределителен сектор - съществува „икономия от мащаба“; дълготрайните активи на подобни фирми се финансират предимно от дълг (стратегически избор на мениджърите); левъриджа увеличава печалбите; фирма, която увеличава своята газоразпределителна инфраструктура се очаква да генерира процентно по-високи печалби в бъдеще, и други. Допълнително направихме някои емпирични допускания за поведението на фирмите от сектора - при ниски цени на газта спрямо електричеството основна стратегия на фирмите е да взимат заеми, за да се разрастват по време на благоприятните условия; ако дадена фирма увеличава процентно доходите си в последователни години, то се очаква тя да спре да се кредитира и да започне да намалява своя дълг; фирмите с много собствен капитал в сравнение с дълга си се очаква да поддържат тази тенденция; и други практики и стратегии. Едно от практичните открития е, че, ако фирма иска да покрие краткосрочните си задължения посредством текущите си активи, тогава тя най-вероятно ще увеличи дълга си за следващия отчетен период. Качеството на намерените модели и смислените изводи, които носят, потвърждават ползата от моделиране на ключови финансови съотношения за даден икономически сектор, както и възможността подобна цел да се реализира на практика.

Глава 4

Регресионна техника за редуциране на влиянието на зависимите предиктори

Много статистически практики, например такива, които често се използват в екологията и макроикономиката, са чувствителни на зависими (корелирани) данни - [Belsley, 1991], [Chatfield, 1995] и други. При наличие на мултиколинеарност някои от оценките за коефициентите на предикторите в модела са нестабилни, оценките на коефициентите им са с по-голяма грешка (за нагледност вижте [Wheeler, 2007]), следователно статистики на този модел са повлияни. При мултиколинеарност влиянието на отделни предиктори не може да бъде отделено едно от друго, което води до модел с колебаещи се оценки на регресионните коефициенти - [Meloun, 1992].

В тази глава представяме техника за машинно самообучение с учител, която чрез преобразуване на предикторите от даден модел в множество от ортогонални помежду си такива и последващо премахване на част от компонентите, цели да минимизира грешките на конкретния модел за външни наблюдения при наличие на мултиколинеарити проблем. Базата за техниката е теоретично доказана. Представени са възможните ѝ приложения. Описани са и примери, използващи наблюдаваните данни за компаниите от газовия сектор на РБ. Обсъдени са сходствата и различията ѝ с двете основни техники за ортогонализиране в регресионния анализ - *Метода на главните компоненти* (Principal component regression) и Регресия по частичните корелации (Partial least squares regression).

4.1 Техника за справяне със зависими предиктори

4.1.1 Теоретично описание на базата за предложената техника

Нека Y е зависимата (моделираната) променлива, а $X_1, X_2, ..., X_k$ са предиктори. Данните са нормализирани по съответното им стандартно отклонение, т.е. центрирани и скалирани до единично стандартно отклонение. Нека предположим, че вече е изчислена множествена регресия и резултата е:

$$\hat{Y} = a_1.X_1 + a_2.X_2 + \cdots + a_n.X_k \tag{4.1}$$

За момента номерацията на X_i, $i = 1, ..., k$ не е съществена. Нека изчислим регресия на X_1 с $X_2, ..., X_k$. Ще запишем регресионното уравнение по следния начин:

$$\hat{X_1} = b_{1,1}.X_2 + b_{2,1}.X_3 + \cdots + b_{k-1,1}.X_k$$

Остатъците от тази регресия ще се бележат с R_1.

$$R_1 = X_1 - [b_{1,1}.X_2 + b_{2,1}.X_3 + \cdots + b_{k-1,1}.X_k]$$

Сега, изчисляваме регресия на X_2 с $X_3, ..., X_k$.

$$\hat{X_2} = b_{2,2}.X_3 + b_{3,2}.X_4 + \cdots + b_{k-1,2}.X_k$$

Съответните остатъци означаваме с R_2.

$$R_2 = X_2 - [b_{2,2}.X_3 + b_{3,2}.X_4 + \cdots + b_{k-1,2}.X_k]$$

Ако продължим процедурата по този начин, ще получим:

$\hat{X_3} = b_{3,3}.X_4 + b_{4,3}.X_5 + \cdots + b_{k-1,3}.X_k$

$\cdots$

$\hat{X_{k-1}} = b_{k-1,k-1}.X_k$

Остатъците на съответните модели ще са $R_3, R_4, ..., R_{k-1}$.

Следващата стъпка на алгоритъма е да изчислим регресия на Y с $R_1, ..., R_{k-1}, X_k$, вектора на предвидените стойности от тази регресия бележим с:

$$\hat{Y_*} = c_1.R_1 + c_2.R_2 + \cdots + c_{k-1}.R_{k-1} + c_k.X_k \tag{4.2}$$

Теорема 4.1. *Оценките* $\hat{Y}$ *от 4.1 и* $\hat{Y}_*$ *от 4.2 съвпадат, т.е. регресията с първоначалните предиктори и регресията с компонентите* $R_1, ..., R_{k-1}, X_k$ *връщат едни и същи прогнозни стойности.*

Теорема 4.2. *:* $R_1, R_2, ..., R_{k-1}, X_k$ *са ортогонални по между си.*

Ортогоналността на $R_1, R_2, ..., R_{k-1}, X_k$ означава, че:

- Корелациите между тях са равни на 0;
- Те не се влияят по между си, т.е., ако моделираме Y с R_i и получим коефициент пред R_i, моделираме Y с R_j и получим коефициент пред R_j, то тези два коефициента ще са същите като съответните им коефициенти при моделиране на Y с R_i и R_j;
- Носят уникална информативност за модела и могат да бъдат разглеждани като отделни модели с по един предиктор, които се явяват отделни независими части на общия модел;
- Ковариационната матрица на коефициентите за регресия с ортогонални предиктори е диагонална матрица.

$R_1, R_2, ..., R_{k-1}, X_k$ са „строителните блокове“ на регресионния модел, те представляват ключовият елемент в тази глава. $R_1, R_2, ..., R_{k-1}, X_k$ ще бъдат наричани накратко **„компоненти“** и ще се означават с $Comp_1, Comp_2, ..., Comp_k$.

4.1.2 Създаване на компоненти

От изложението на метода стана ясно, че в зависимост от реда на избор на предиктори се образуват различни компоненти. Ако n е броя на предикторите, то това означава, че съществуват $n!$ различни начина да получим множество от компоненти. От всички множества от компоненти от особен интерес са тези, които съдържат компоненти, които ще са статистически незначими или допринасят с много малко към R^2 на модела. Идеята е впоследствие те да бъдат премахнати и по този начин евентуално да се редуцира грешката. Ще се предложат няколко начина за генериране на подходящи компоненти:

- **Изчерпващ подход** - поетапно се пресмятат всички възможни варианти за генериране на компоненти. Като се запазват първите няколко оптимални модела. Имайки предвид, че при 10 предиктора имаме 3 628 800 варианта, а при 11 вече са 11 пъти повече, то този подход е удачен за сравнително малко предиктори в модела. Ако се вземе предвид изискването за простота на моделите, то подходът е удачен в повечето случаи, защото гарантира, че няма да се пропусне оптималния

резултат. Самото сравняване на оценените моделите от компонентите е чрез кросвалидация [Davison and Hinkley, 1997, Hawkins et al., 2003, Mevik and Cederkvist, 2004];

- **Евристични подходи** - подобни подходи се използват, когато множеството от възможни колекции от компоненти е твърде голямо, т.е. при модели с повече от дузина предиктора. Ще предложим няколко потенциално добри метода за построяване на компонентите. При поетапното построяване на всяка стъпка се избира: предиктора с най-висок VIF; предиктора с най-висока по модул t-стойност; компонентата, която има най-високо R^2 спрямо моделираната променлива; възможни са много други подходи. Добра стратегия е прилагането на множество евристични подходи и избиране на оптималния техен резултат.

4.2 Алгоритъм на техниката в случая на евристичен подход чрез t-стойности

За да бъдем напълно ясни как точно функционира представената техника, ще опишем чрез алгоритъм един от предложените подходи от Раздел 4.1.2:

1. Прогнозираме Y с предиктори $X1$, $X2$,..., Xk. Всички дефинирани обекти са вектор стълбове от числа;
2. Изчисляваме множествена регресия Y срещу $X1$, $X2$, ... , Xk;
3. Изчисляваме t-стойностите ti, които съответстват на съответните коефициенти от регресията пред Xi, $i = 1, ..., k$ (избираме само от първоначалните предиктори). ti представлява съотношението оценка на коефициент спрямо стандартната грешка за тази оценка;
4. Избираме предиктор Xi, който съответства на най-голямото по абсолютна стойност ti (друга евристика е да избираме Xi с най-малкото ti);
5. Изчисляваме множествена регресия Xi срещу $X1$, $X2$, ...,$X(i-1)$, $X(i+1)$, ... , Xk. Записваме в Ei остатъците от тази регресия;
6. Формираме нов предиктор $Ci = Ei$; Заместваме Xi с Ci и се връщаме обратно на Стъпка 2;
7. Край - алгоритъмът приключва, когато е останал само един единствен не променен предиктор.

В последствие строим регресия Y срещу получените компоненти и премахваме най-лошо оценените по p-стойност от тях, ако премахването на компонента води до чувствително намаляване на R^2 на модела или друга приоритетна характеристика, то премахването на компонентата по-скоро ще влоши модела.

Следва да се отбележи, че останалите евристични подходи следват аналогичен алгоритъм, но с различен критерий за избор на предиктор за формиране на следващата компонента.

4.3 Обобщение

В тази глава изложихме техника за машинно самообучение с учител, която посредством ортогонализация на предикторите на множествена регресия и манипулации върху част от тях, при наличие на мултиколинеарити, цели по-малки грешки при предвиждане на нови наблюдения. Теоретично обосновахме базата за предложената техника. Ако n е броя на предикторите, то $n!$ е броя на възможните множества от по n ортогонални предиктора, които техниката строи. За всяко едно такова множество от ортогонални предиктори ние показахме, че модел с тях едно към едно отговаря на първоначалния регресионен модел. Изведохме и ковариационна матрица, която описва връзката на коефициентите на модела с ортогонални променливи с ковариационната матрица на първоначалния модел. От изведената ковариационна матрица заключихме, че е възможно да повлияем на тази матрица посредством премахване или комбиниране на ортогоналните предиктори. Премахването на някой от ортогоналните предиктори не винаги е води до намаляване на грешката в оценката на модела. Тогава може да търсим в друго множество от ортогонални предиктори. Самото търсене сред множества от предиктори може да става по избран критерий или с пълно изчерпване на възможните случаи.

Глава 5

Крос-валидирана постъпково конструирана множествена регресия

тази глава ще бъде изложена техника за машинно самообучение с учител базирана на множествена регресия, която модифицира регресионните коефициенти, групирайки ги в компоненти, чрез постъпкова минимизация на коренуваната средноквадратична грешка на модела, получена при поелементно крос-валидиране. Техниката е дефиниране за два и повече предиктора. Коментирани са възможните ѝ приложения и при какви условия се очаква да е подходящ заместител на стандартната множествена регресия. Техниката е демонстрирана върху ROA модела на фирмите от газовия сектор.

5.1 Същност на метода

Техниката за машинно самообучение с учител, която предлагаме, обединява две по две избрани предиктори в компоненти, намалявайки по този начин броя на независимите променливи. Същността на метода е начина, по който се извършва обединяването на предиктори в компонента. Всяка компонента се извежда чрез обединяване на два предиктора или вече получени компоненти. Нека за простота приемем, че сме оценили множествена регресия без свободен член с два предиктора и резултата е:

$$Y = a_1 X_1 + a_2 X_2 + e. \tag{5.1}$$

Комбинираме предиктори X_1 и X_2 в компонентата Z. Допълнително, нека X_1 и X_2 са индексирани така, че оценката a_1 с очаквана грешка σ_1 на коефициента пред X_1 е по-голяма по абсолютна стойност от оценката a_2 с очаквана грешка σ_2 на коефициента пред X_2. При тази индексация по конструкция се очаква, че параметъра k,

когото търсим ще е малко число, но големи стойности не пречат на техниката. Ще дефинираме $Z(k) := X_1 + k * X_2$, където търсим оптималното k, докато минимизираме *крос-валидираната коренувана средноквадратична грешка* $RMSECV(k)$ за модел с компонентата $Z(k)$, която замества X_1 и X_2 (за модел с повече от два предиктори с $Z(k)$ добавяме и останалите), вижте Уравнение 5.2.

$$\min_k RMSECV(k) = \min_k \sqrt{\frac{\sum_{i=1}^{n} \epsilon_{i,k}^2}{n}}, \quad k \in (\frac{a_2 - \mu\sigma_2}{a_1 + \mu\sigma_1}, \frac{a_2 + \mu\sigma_2}{a_1 - \mu\sigma_1}),\ \mu > 0, a_1 \neq \pm\mu\sigma_1, \tag{5.2}$$

където $\epsilon_{i,k} = Y(i) - f_i(k)$ е грешката за i-тото наблюдение от модел $f_i(k)$. $f_i(k)$ е модела с компонентата $Z(k)$ вместо двата предиктора X_1 и X_2, оценен от цялото множество наблюдения без i-тото наблюдение, с функционална форма като в (5.3). Търсим k в интервал около стойността $k_0 = \frac{a_2}{a_1}$, $k_0 \in [-1, 1]$ по построение, защото при $k = k_0$ компонентата $Z(k_0)$ ще е такава линейна комбинация на X_1 и X_2, която след оценка с множествена регресия и разписването ѝ чрез X_1 и X_2 ще върне първоначалния регресионен модел от (**??**).Идеята за интервала $(\frac{a_2-\mu\sigma_2}{a_1+\mu\sigma_1}, \frac{a_2+\mu\sigma_2}{a_1-\mu\sigma_1})$ идва от това, че търсим стойности за k в околност на k_0, а самото k_0 е съотношение на оценките на двата коефициента a_1 и a_2, т.е. отклонението от k_0 ще представим като функция на стандартни грешки съответно σ_1 и σ_2 на двата коефициента a_1 и a_2.

За произволно k модела от (5.1) се преобразува в:

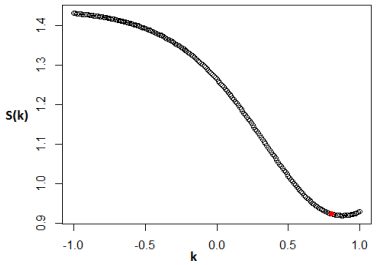
$$Y = a(k)Z(k) + e(k), \tag{5.3}$$

където $Z(k) := X_1 + k * X_2$, а $a(k)$ е регресионен коефициент, оценен след получаване на компонентата $Z(k)$, а $e(k)$ е вектора на остатъците. Всеки модел от вида (5.3) може лесно да се изрази с предиктори X_1 и X_2:

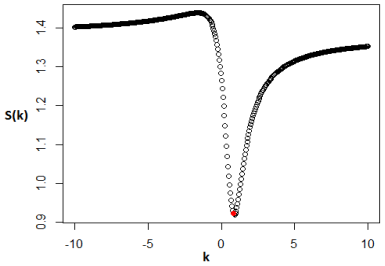
$$Y = a(k)Z(k) + e(k) = a(k)(X_1 + k * X_2) + e(k) = a(k)X_1 + (a(k)k)X_2,$$

това запазва възможността за интерепретиране на модел, получен по описаната процедура.

Примерно графичното онагледяване на $RMSECV(k)$ може да се види на Фигура 5.2. С червено е отбелязана точката k_0 на нормално оценената регресия, вижда се, че тази точка е близо до минимума, но съществува друга точка k_*, която минимизира функцията. Сега, ако построим компонентата $Z(k_*)$, то регресията със $Z(k_*)$ има по-малка стойност на RMSECV от първоначалния модел, т.е. очаква се този модел при преизчисляване с различни наблюдения и повторно използване да има по-ниска грешка от множествената регресия с двата предиктора X_1 и X_2.



Фигура 5.1: Функцията $RMSECV(k), k \in [-1, 1]$ от примера.



Фигура 5.2: Функцията $RMSECV(k), k \in [-10, 10]$ от примера.

5.2 Алгоритмично представяне на „ядрото“ на техниката

Ядрото на техниката се състои в комбинирането на два избрани предиктора в компонента. Как избираме кои да са тези два предиктора предлагаме в следващия раздел. Ще реализираме алгоритмично функция **Core(Y,X,p,q)**.

Вход: **Y** – вектор от n стойности на моделираната променлива; **X** – матрица n x m, съдържа по n стойности на предикторите X1,. . . ,Xm; цели числа **p** и **q**, $p, q \in [1, \ldots, m]$, които да укажат, кои предиктори от X ще манипулираме.

Изход: **k*** - оптималното k за изготвяне на компонентата $Z = Xp + k.Xq$; **p**, **q** - отново ги връщаме, защото може да са разменени.

1. По метода на най-малките квадрати пресмятаме коефициенти $\alpha_0, \alpha_1, ..., \alpha_m$ за множествена регресия $Y = \alpha_0 + \alpha_1 X_1 + ... + \alpha_m X_m$;
2. Избираме си ориентировъчно $k0 = \frac{\alpha_q}{\alpha_p}$, ако $abs(\alpha_q) < abs(\alpha_p)$, в противен случай разменяме стойностите на p и q и повтаряме стъпката;
3. Интересуваме се да построим компонента $Z = Xq + k.Xp$, неизвестното k приема стойности в интервал около $k0$. За да построим този интервал ни трябват стандартните грешки σ_1 и σ_2 съответно на оценките α_q и α_p. Те съответстват на коренуваните q-ти и p-ти елемент от главния диагонал на ковариационната матрица на коефициентите;
4. Интервалът, където търсим оптималното k е $k \in (\frac{a_2 - \mu\sigma_2}{a_1 + \mu\sigma_1}, \frac{a_2 + \mu\sigma_2}{a_1 - \mu\sigma_1})$, $\mu > 0, a_1 \neq \pm\mu\sigma_1$. Полагаме $\mu = 4$. Така дефинираният интервал минимизира възможността да се пропусне глобалния минимум при търсене с фиксирана стъпка, защото е скалиран спрямо големината на коефициентите α_q и α_p, и съответните им стандартни грешки;
5. За всяка стойност на k, която ще вземем от дефинирания интервал имаме следната процедура за RMSECV:

(а) По МНК оценяваме множествена регресия Y срещу предиктори $X_1, X_2, ..., X_{p-1}, X_{p+1}, ..., Xq-1, X_{q+1}, ..., X_m, Z(k)$;

(б) Сега може ускорено да намерим грешките след 1-1 крос-валидация ϵ_i, $i = 1, ..., n$ посредством формулата от Теорема 5.1;

(в) На база на ϵ_i, $i = 1, ..., n$ пресмятаме и *Коренуваната средноквадратична грешка след крос-валидация*, вижте Формула RMSECV.

6. Стойност на k, която ще минимизира RMSECV може да се търси с произволен branch&bound алгоритъм, но за пълнота ще дадем собствена процедура:

(а) Дефинираният интервал за стойности на k разделяме на 100 равни части, т.е. в моментно разполагаме със 100 стойности за k;

(б) За всяка от получените стойности за k, без първата и последната, намираме RMSECV и ги записваме като пазим кое RMSECV на кое k съответства;

(в) Избираме това $k = k*$, което има минимална RMSECV. Презаписваме интервала, където търсим оптималното k, като за лява граница използваме съседното от ляво k на $k*$, а за дясна граница съседното от дясно.

(г) Новополучения интервал отново подразделяме на 100 равни части и повтаряме стъпките;

(д) Край - цикълът спира, когато разликата между предходното минимално RMSECV и подобреното такова е много малка, например разлика под $\frac{2}{10^9}$.

7. Край - когато се достигне до задоволително $k*$.

5.3 Методът при повече от два предиктора

За повече от два предиктора ще предложим следната процедура:

1. Моделираме Y с множествена регресия чрез предиктори X1,. . . ,Xm записани като стълбове в матрица X;
2. Избираме най-лошо оценения коефициент на предиктор по съответното му p-число. След това намираме предиктор, който е най-силно корелиран с него (по коефициента на Пирсън), нека избраните два предиктора да отговарят на индекси p и q. Тогава комбинираме избраните предиктори в една компонента чрез функцията Core(Y,X,p,q), която представихме в Раздел 5.2;
3. Изхода от функцията Core(Y,X,p,q) са k*, p*, q*. Презаписваме $p = p*$, $q = q*$, създаваме компонента $Zpq = Xp + k * Xq$
4. Връщаме се в Стъпка 1) като в матрицата X за X1,. . . ,Xm предиктори Xp и Xq са заменени с компонентата Zpq;

5. Край - алгоритъмът спира, когато за всички предиктори постигнем p-стойности над определен лимит или когато броя на предикторите достигнем определен минимум. Предлагаме край на процедурата да настъпва, когато t-стойностите са над 4.5 по абсолютна стойност (съответства на p-стойност около 0.000035) или когато броят на получените предиктори е около $\frac{n}{2}$;

Получените компоненти и неизменените предиктори се използват вместо n-те първоначални предиктори.

Теорема 5.1. *Грешките на външните наблюдения ϵ_i, $i = 1, ..., n$, използвани в RMSECV(k) функцията, могат да бъдат изчислени от следната формула:*

$$\epsilon_i = \frac{\hat{\xi}_i}{1 - h_{i,i}}, \; i = 1, ..., n, \tag{5.4}$$

*където $\hat{\xi}_i$ е грешката на i-тото наблюдение на първоначалния регресионен модел, n е броят на наблюденията, а $h_{i,i}$ е i-ят елемент на диагонала на матрицата $H = X(X^tX)^{-1}X^t$ (*hat matrix*), където X е матрицата от предикторни променливи, X^t е съответната транспонирана матрица.*

5.4 Обобщение

В тази глава представихме техника за машинно самообучение с учител, която цели да оцени линеен модел, така че той да е по-робастен (robust) към нови наблюдения в множеството, от което се оценява, по отношение на представянето му при прогнозиране на наблюдения извън това множество. С други думи, техниката крос-валидира самите оценки на регресионните коефициенти, което води до по-точна оценка на база на наличните наблюдения. Тестовете на техниката върху примера показват сходни прогнозни резултати с тези на множествена регресия, но при условие, че новата техника е редуцирала предикторите до 3 (при 7 за регресията) и ги е оценила само с 38% от наличните данни (не най-високо информативните 38% от множеството). Допълнително, техниката изчислява компоненти, които са чувствително по-малко от първоначалния брой предиктори и тези компоненти са изчислени така, че да дават колкото се може по-ниски грешки при повторно оценяване на коефициентите пред тях с допълнителни наблюдения. Това носи ползи, когато моделите се преизчисляват в реално време с голямо количество наблюдения, защото ще се преизчисляват в пъти по-малко коефициенти. Постигнатите резултати и уникалността на техниката (подобен подходен не е открит в литературата) стимулират предстоящи изследвания по отношение на

завършена теоретична обоснованост и сферата на практическото приложение на метода. Налични са признаци, че модификации на предложения подход ще допринасят в сферите на класификационните методи и времевите редове.

Глава 6

Използване на резултатите при оценка на риска

В тази глава глава ще представим базова рамка за оценка на риска от фалит на фирмите от газовия сектор в България на база на разработките от Глави 2, 3, 4 и 5. Ще представим концепция как да използваме ROA и FL финансовите съотношения заедно. Ще покажем как да третираме проблемни входни данни за моделите, как да подобряваме резултатите от моделите за конкретна фирма и как да получим допълнителна информация за бъдещите прогнози. В края на главата ще представим и ориентировъчни критерий, по които да определяме дали една фирма е добре финансово или е нестабилна, т.е. критерий за оценка на риска от фалит.

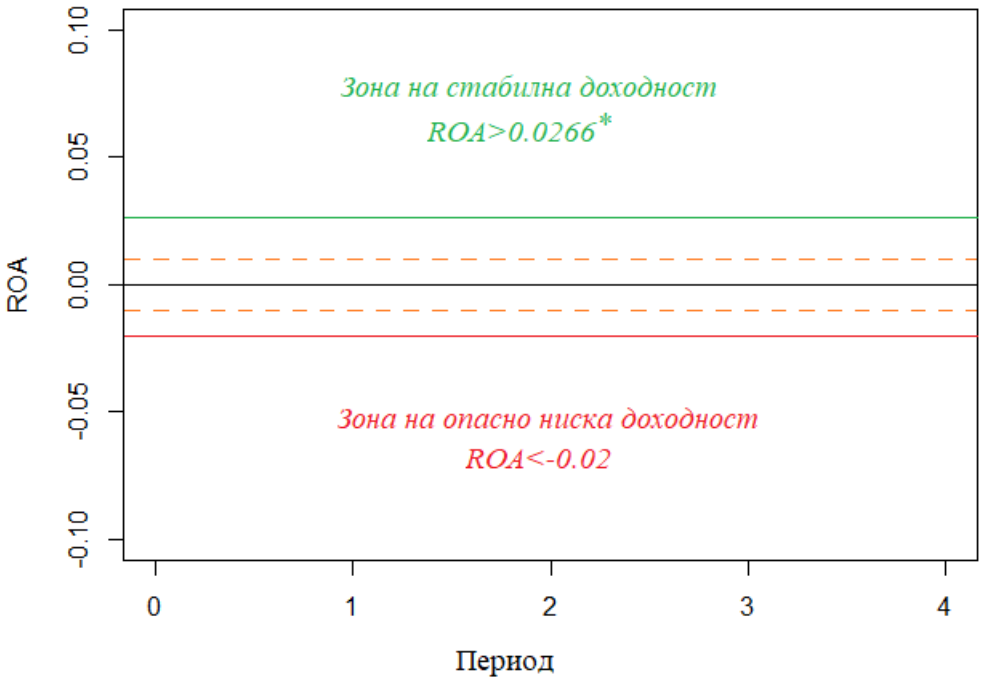
6.1 Оценка на риска при фирмите от газоразпределителния сектор

Поради твърде оскъдна информация за фалирали фирми, директно моделиране на риска от фалит на фирми от газоразпределителния сектор в България е невъзможна. Но това не означава, че не могат да се изготвят ориентири. На база на:

- ROA и FL моделите;
- наблюденията по години на средните стойности за печалбите и задлъжнялостта;
- опита от анализа на изпадналата в неплатежоспособност фирма „ГАЗТРЕЙД-СЛИВЕН" ЕООД;
- *Течния модел* на Бийвър, наблюденията му върху отделни ФС по отношение на риска от фалит и моделите за оценка на риска, които изложихме в уводната глава;

ще предложим базов похват за качествена оценка на здравето на дадена фирма.

Нека да разгледаме първо съотношението ROA. Следва да се определи и праг, над когото дадена фирма е финансово здрава. Доходността на фирмите спрямо стойностите на ROA обобщаваме на Фигура 6.1. Забележете, че стойността за прага на зелената

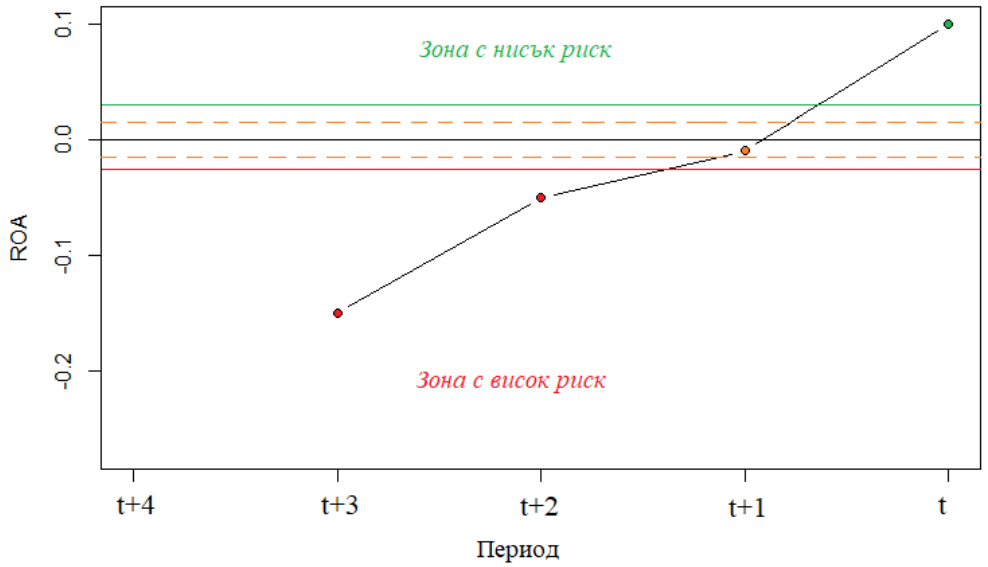


Фигура 6.1: Различни нива на стойностите за ROA.

зона е със звездичка, с което индикираме, че той е условен и зависи от цената на газта.

Как да използваме съответната схема за нивата на ROA, ще дадем с пример. На Фигура 6.2 показваме движението на печалбите на една хипотетична фирма. Вижда се, че с ускорително темпо печалбите намаляват. В продължение на три години те са отрицателни. Това, което предлагаме, е, че фирма, която задържа ROA под червената граница в продължение на повече от две години може да фалира на третата година. Ускорителните темпове на отрицателните печалби са допълнителен сигнал за подобна тенденция. Разбира се, следва да видим какво се случва със задлъжнялостта, за да бъдем по-сигурни в предположенията си.

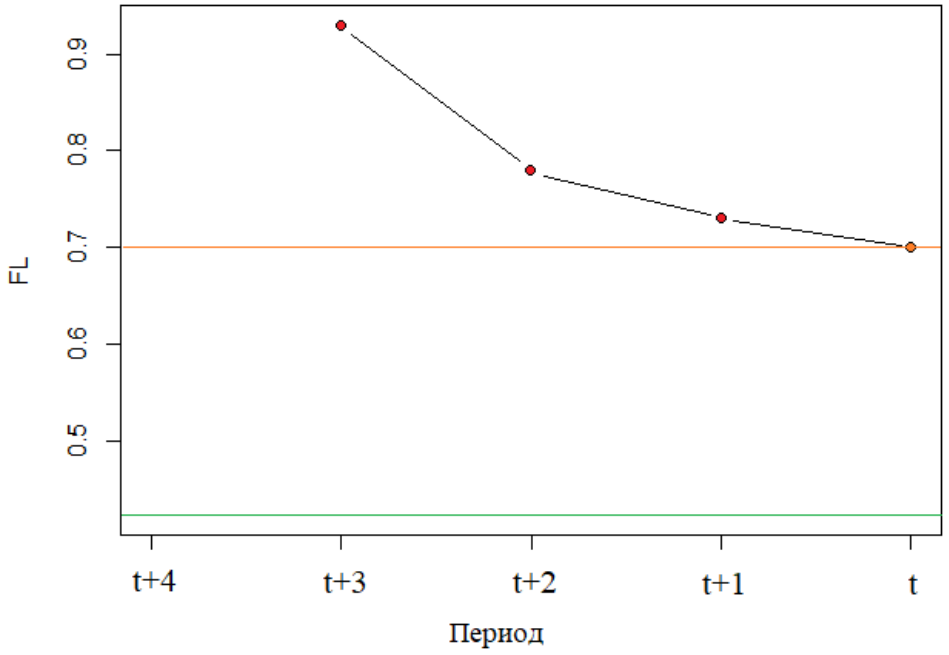
На Фигура 6.3 сме изложили предложение за подразделяне на стойностите на задлъжнялостта FL. Разграничават се няколко нива на задлъжнялост. Отвъд оранжевата зона всяка една фирма се третира като потенциално уязвима фирма. Границата 1 е разделната черта за екстремна задлъжнялост, т.е. над 100%. В множеството от данни са налични примери за фирми с екстремна задлъжнялост, те или фалират, или са новообразувани такива, зад които стои дъщерна компания гарант. Нека да продължим хипотетичния пример за фалираща фирма, когото представихме със стойности



Фигура 6.2: Хипотетичен пример за потенциално, фалираща фирма в близко бъдеще спрямо ROA.

Фигура 6.3: Различни нива на стойностите за FL.

за ROA. На Фигура 6.4 се вижда как за три годишен период хипотетичната фирма увеличава задлъжнялостта си ускорително, като още в началния си етап е била със задлъжнялост около 70%. При такива тенденции за ROA и FL очакваме тази фирма



Фигура 6.4: Хипотетичен пример за потенциално, фалираща фирма в близко бъдеще спрямо FL.

да фалира в близък период от време.

6.2 Обобщение

В тази глава, използвайки директен пример, показахме как разработките в този дисертационен труд могат да бъдат имплементирани в процеса на оценка на риска от фалит за фирмите от газоразпределителния сектор на България. Постъпково изложихме как да оценяваме ROA и FL за бъдещи периоди. В тези стъпки включихме третирането на проблемни наблюдения, взимането предвид индивидуални фирмени особености, манипулирани данни, доверителни интервали и други тънкости. На база на събраното множеството от данни, фалиралата фирма и изготвените модели, оформихме и критерий за ROA и FL, които да служат за ориентир при оценка на здравето на произволна фирма от газоразпределителния сектор в България.

Глава 7

Заключение

В този дисертационен труд бе извършена работа, фокусирана върху компютърна симулация и оценка на фирмени рискове. Обект на това изследване бе газоразпределителния сектор в РБ. В уводната глава въведохме читателя в сферата на оценка на рисковете, необходимите термини от финансите, формулирахме проблема, подбрахме подходящи статистически методи за моделиране и проверихме за сходни разработки. Основни резултати от главата са, че всички рискове могат да се сведат само до няколко и най-вече до риска от фалит, че финансовите съотношения ROA и FL са ключови при оценка на риска от фалит, че множествената регресия е подходящия статистически инструмент за моделиране и че няма подобно изследване за газоразпределителния сектор в България и изобщо. Техниките и процедурите за боравене с данни и регресионни модели, които сме използвали, описахме в Глава 2. Съставената методология проследява процеса от събирането и съхранението на данните (данните са събрани специално за това изследване) до постъпковото моделиране и начините за боравене с готовите модели. Основния резултат от главата е, че създадохме полуавтоматизирана процедура за строене на регресионни модели. За изчислителната работа, графиките и симулациите използваме езика за статистическа обработка на данните R. В глава 3 направихме кратък обзор на газоразпределителния сектор в България и следвайки изготвената методологията от Глава 2, изчислихме и верифицирахме модели за печалбите и задлъжнялостта. Рядко срещана характеристика на получените модели е тяхното високо R^2 и едновременното присъствие на финансови съотношения и макроикономически величини. Моделите могат на база на данните от настоящата година да прогнозират година напред във времето, а посредством коефициентите от моделите направихме редица изводи за поведението на фирмите в сектора. Тези изводи пряко или косвено засягат оценка на рисковете и хвърлят светлина за някои специфики в сектора. В процеса на оценка на регресионни модели чрез езика R се стигна

до съставянето на две техники за машинно самообучение с учител, които в определени ситуации да могат да се използват преимуществено пред множествената регресия за прогнозиране на нови наблюдения. В Глава 4 е предложена ортогонализационна техника, която в условия на високи корелации между предикторите в даден модел да може да го оцени по такъв начин, че да намали грешката на множествената регресия. Основната идея зад техниката е теоретично обоснована в същата глава. Предлагаме и няколко алгоритъма за прилагането ѝ. В Глава 5 описваме и втората техника, която наричаме *Крос-валидирана постъпково конструирана множествена регресия*. Тази техника, посредством крос-валидация, така оценява коефициентите на даден линеен модел, че той да се представя колкото се може по-добре при прогнозиране на бъдещи наблюдения, дори след повторна оценка на съответния модел с нови наблюдения. Техниката е особено полезна, когато разполагаме с малък брой наблюдения за оценка на модела. Алгоритъмът, по който се оценява тази регресия е създаден така, че да намалява ефекта от корелации между предикторите. Последната глава се явява своеобразно финализиране на поставените цели в дисертационния труд. Излага се техника за съвместно използване на ROA и FL моделите за оценка на финансовото състояние на фирма от газоразпределителния сектор. Представят се критерий за стойностите на ROA и FL и взаимовръзката им с риска от фалит. Представяме и конкретен пример за фалирала фирма и как да се справим с трудностите при прилагане на моделите.

7.1 Научни и приложни приноси

Проектът с газоразпределителните фирми в България е базиран на реални данни, целият процес на постигане на целите е успешно реализиран. В процеса на решаване на задачата се разрешават необходимото множество реални практически трудности, от сферата на информационното осигуряване до прилагането на редица статистически инструменти и програмен код. Получените прогнозни модели са с високо, за икономическите сфери, ниво на точност. Избраните променливи, участващи в моделите, са икономически и финансово достоверни и представляват база за извеждане на полезни наблюдения върху финансовото представяне на газ фирмите. Подобна разработка в сферата на газоразпределителните фирми не е предприемана.

Методологията, развита за конкретно избрания сектор, поставя основите на технологична рамка, приложима за произволна група от фирми или цял отрасъл, която може да се използва като база за изработването на система за взимането на решения, моделно базирана, помагаща в процеса на управление на рискове, икономически анализи, финансови анализи и други.

В процеса на работа са разработени и представени две техники за машинно самообучение с учител, които оценяват коефициентите на множествена регресия. Целта е подобряване на прогнозната точност, т.е. намаляване на грешките при прогнозиране на нови за модела наблюдения, в определени случаи - съответно мултиколинеарити и малък брой наблюдения. Базата за новите техники подлежи на теоретична обосновка. Те могат да бъдат използвани във всички проблеми, където се използва и множествената регресия.

Открояват се следните новости:

- Създадена е техника за машинно самообучение с учител, ортогонализираща предикторите на едно регресионно уравнение при наличие на силна корелираност между тях. Ортогонализираните предиктори са построени така, че имат удобна трактовка за разлика от други ортогонализиращи методи, например метода на главните компоненти. Базата за построяване на компонентите е теоретично обоснована. При определени условия премахването или манипулирането на част от компонентите повлиява така на коефициентите на линейния модел, че намалява грешката от модела. Предложихме редица разновидности на тази техника;
- Създадена е техника за машинно самообучение с учител, минимизираща среднаквадратичната грешка след крос-валидация на множествена регресия при наличие на сравнително малък брой наблюдения за оценка на модела и евентуален мултиколинеарити проблем. Резултатът е регресионно уравнение с по-малка грешка при прогнозиране на бъдещи наблюдения, поради поетапно крос-валидиране на коефициенти в модела.
- В този труд се описва първото изследване върху газоразпределителните фирми на България през призмата на финансови съотношения и макроикономически величини по отношение на оценка на рисковете. Изследването е структурирано така, че да служи като база за *система за взимането на решения* по отношение на оценка на риска в сектора. Допълнително, разработената функционалност може да се модифицира за други икономически сектори. В изследването се открояват следните особености:

  – Отчитане на счетоводни, отраслови и макроикономически фактори - при образуване на моделите е търсено измежду над 80 променливи, в направения прочит на литературата не е наблюдавано изследване с повече от 30 такива. В предишни изследвания макроикономическите величини се пренебрегват, докато регресионните модели разработени в следващите глави показват тяхната значимост. Дефинират се и се анализират специфични за конкретния сектор величини, някои от които включваме в крайните модели;

– Математическа прецизност при прилагане на статистическите модели - обикновено статиите в икономическата литература пренебрегват в една или друга степен теоретичните условия за прилагане на различни статистически подходи, в това число множествена регресия, при която без съблюдаване на условията на Гаус-Марков нищо не гарантира за точността ѝ. В дисертацията се предлага полуавтоматизирана процедура за конструиране на регресионни модели;
– Предлагаме цялостна методология за начална оценка на риска от фалит в сектора. Основна база са две високоинформативни ФП, измерители на производителността и задлъжнялостта - съответно ROA и FL. Методологията включва конструирането на два добре оценени модела за тези ФП, изводи направени на база на тези модели, иновативна процедура за третиране на проблемни входни данни на база на логистичната функция, процедура за съвместно използване на моделите, и критерий за стойностите на тези ФП за оценка на финансовото състояние на дадена компания.

7.2 Апробация на резултатите

Научните резултати са докладвани на следните научни конференции:

- 12th Annual Meeting of the Bulgarian Section of SIAM, 20 - 22 декември, 2017, София, България. Доклад на тема "Cross-validated sequentially constructed multiple regression";
- Трета годишна докторантска конференция на НБУ, 09 – 11 февруари, 2018 година, местност Бачиново, Благоевград. Доклад на тема „Регресионен анализ - добри практики и особености“;
- 14th Annual International Conference on Computer Science and Education in Computer Science, лятото на 2018, Бостън, САЩ. Доклад на тема "Statistical methods for the course AD 699 Data Mining for Business Analytics (Boston University)";
- 13th Annual International Conference on Computer Science and Education in Computer Science, лятото на 2017, Албена, България. Доклад на тема "Prediction error in multiple regression model";
- Втора годишна докторантска конференция на НБУ, 24 – 26 февруари, 2017 година, местност Бачиново, Благоевград. Доклад на тема „Класически и съвременни методи за моделиране на електропотреблението“;
- Ежеседмичен семинар на департамен Информатика, 2017. Доклад на тема „Дългосрочно прогнозиране на крайното електропотребление в България до 2035 година“;

- 12th Annual International Conference on Computer Science and Education in Computer Science, 1 - 4 юли, 2016, Фулда и Нюрнберг, Германия. Доклад на тема "Modified regression technique to reduce effects of multicollinearity";
- Doctoral Conference in Mathematics and Informatics, 15 - 18 октомври, 2015, ИМИ БАН, София. Доклад на тема Modelling company's performance based on financial ratios and macroeconomic variables;
- Първа годишна докторантска конференция на НБУ, февруари 2016 година, местност Бачиново, Благоевград. Доклад на тема „Моделиране представянето на фирма, базирано на финансови съотношения и макроикономически величини“;
- 19th European Young Statisticians Meeting, 1 - 4 септември, 2015, Прага, Чехия. Доклад на тема "Modelling company's performance based on financial ratios";
- Доклади пред експерти на ЦАУР - през периода на обучение докторантът работи в *Лаборатоия за оценка на риска* към НБУ и ЦАУР, където участва в редица проекти. Проектите са основно в сферата на енергетиката - природен газ и електропотреблението. Ключови акценти са газов хъб „Балкан“; АЕЦ „Белене“; дългосрочно прогнозиране на електропотреблението, поръчка на българския „Електроенергиен системен оператор“ (ECO).

Публикувани са четири научни статии по темата на дисертацията:

- Angelov, S. and Stoimenova, E.(2016). Modified Regression Technique to Reduce Effects of Multicollinearity. *Computer Science and Education in Computer Science*, 12: 353-367;
- Angelov, S. and Stoimenova, E.(2017). Prediction error in multiple regression model. *Computer Science and Education in Computer Science*, 13;
- Ангелов, С. (2018). Класически и съвременни методи за моделиране на електропотреблението. *Език и Публичност* (докторантски брой), 2:87-101;
- Angelov, S. and Stoimenova, E.(2019). Cross-Validated Sequentially Constructed Multiple Regression. *Advanced Computing in Industrial Mathematics, Springer*, pp 13-22, SJR.

Библиография

AIRMIC, Alarm, I. (2010). A structured approach to enterprise risk management and the requirements of iso 31000. https://www.theirm.org/media/886062/ISO3100_doc.pdf.

Belsley, D. (1991). Conditioning diagnostics: collinearity and weak data regression. *Wiley Online library. D.A.Belsley*, 8:343–348.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of Royal Staistical Society*, 158:419–466.

Davison, A. and Hinkley, D. (1997). Bootstrap methods and their application. cambridge series in statistical and probabilistic mathematics. *Cambridge University Press*.

Hawkins, D., Basak, S., and Mills, D. (2003). Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.*, 43(2):579—-586.

Meloun, . (1992). Crucial problems in regression modelling and their solutions. *Analyst*, 127:433–450.

Mevik, B. and Cederkvist, H. (2004). Mean squared error of prediction (msep) estimates for principal component regression (pcr) and partial least squares regression (plsr). *Journal of Chemometrics*, 18:422–429.

Wheeler, D. (2007). Diagnostic tools and a remedial method for collinearity in geographically weighted regression. *Environment and Planning A*, 39:2464–2481.

NEW BULGARIAN UNIVERSITY

Abstract

Faculty of Mathematics and Informatics
Department "Informatics"

Computer simulation and assessment of firm risks

from Slav Emilov Angelov

The aim of the thesis is to model the uncertainty around the financial risk of the firms from a specific economic sector. The subject of the research is the firms from the Bulgarian gas distribution sector. The realisation of the goals is achieved by regarding factors in the spheres of accounting, finance, macroeconomics, probability and statistics. We have created a fully functional conceptual framework during the research which can be implemented in a model-driven decision support system for assessing the risk. Additionally, two machine learning techniques for regression modeling were created. One of them cross-validates the estimates of the coefficients in the model in a step-by-step procedure, while the second handles models with highly correlated model variables by extracting orthogonal components from the predictors. For the realization of the presented techniques and models the language for statistical programming R is used.