ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Кластерный анализ позволяет разделить выборку на группы однородности, однако не дает ни правил, ни четких критериев оценки качества классификации. Поэтому, результаты классификации могут быть неоднозначными и зависеть от искусства пользователя.

Дискриминантный анализ (*ДА*) лишен перечисленных недостатков и включает статистические методы классификации многомерных наблюдений в ситуации, когда исследователь обладает так называемыми обучающими выборками. Этот вид анализа является многомерным, так как использует несколько признаков объекта, число которых может быть сколь угодно большим.

Цель *ДА* состоит в том, чтобы на основе измерения различных характеристик (признаков) объекта классифицировать его, т. е. отнести к одной из нескольких заданных групп (классов) некоторым оптимальным способом. При этом предполагается, что исходные данные наряду с признаками объектов содержат категориальную (группирующую) переменную, которая определяет принадлежность объекта к той или иной группе. Поэтому в *ДА* предусмотрена проверка непротиворечивости классификации, проведенной методом с исходной эмпирической классификацией. Под оптимальным способом понимается либо минимум математического ожидания потерь, либо минимум вероятности ложной классификации.

В общем случае задача различения (дискриминации) формулируется следующим образом. Пусть результатом наблюдения над объектом является построение *k-мерного* случайного вектора $X = (X_1, X_2, \ldots, X_K)$, где $X_1, X_2, \ldots, X_K$ – признаки объекта. Требуется установить правило, согласно которому по значениям координат вектора X объект относят к одной из возможных совокупностей π_i, $i = 1, 2, \ldots, n$.

Правило дискриминации выбирается в соответствии с определенным принципом оптимальности на основе априорной (до опыта, в нашем случае до проведения дискриминантного анализа) информации о вероятностях p_i извлечения объекта из π_i. При этом следует учитывать размер потерь от неправильной дискриминации. Априорные вероятности p_i могут быть либо заданы, либо нет. Очевидно, что рекомендации будут тем точнее, чем полнее исходная информация.

Обычно в задаче различения переходят от вектора признаков, характеризующих объект, к линейной функции от них – к дискриминантной функции-гиперплоскости, наилучшим образом разделяющей совокупность выборочных точек.

Методы дискриминации можно условно разделить на параметрические и непараметрические.

В параметрических известно, что распределение векторов признаков в каждой совокупности нормально, но нет информации о параметрах этих распределений. Здесь естественно в дискриминантной функции заменить неизвестные параметры распределения их наилучшими оценками, произведенными на основе выборочных точек. Правило дискриминации можно основывать на отношении правдоподобия.

Непараметрические методы дискриминации не требуют знаний о точном функциональном виде распределений и позволяют решать задачи дискриминации на основе незначительной априорной информации о совокупностях, что особенно ценно для практических применений.

Таким образом, параметрический (классический) *ДА* применяется при выполнении ряда предположений:

– предположения о том, что наблюдаемые величины непрерывные, измерены как минимум в интервальной шкале и имеют нормальное распределение. Это предположение следует проверять. В модуле программы STATISTICA имеются специальные опции, позволяющие быстро построить гистограммы и нормальные вероятностные графики. Умеренные отклонения от этого предположения допустимы;

– предположения об однородности дисперсий и ковариаций наблюдаемых переменных в разных классах. Умеренные отклонения от этого предположения также допустимы.

Модели, реализованные в методе, являются линейными, а функции классификации и дискриминантные функции – линейными комбинациями наблюдаемых величин.

Дискриминантный анализ базы данных российских предприятий с помощью модуля Discriminant Analysis

Следующий этап исследования это найти правило, согласно которому по значениям 27 финансовых показателей предприятия относят к той или иной группе.

В программе Statistica 6 имеется модуль **Discriminant Analysis** (дискриминантный анализ) – это обучающая система и очень полезный инструмент для поиска переменных, позволяющих относить наблюдаемые объекты в одну или несколько реально наблюдаемых групп; классификации наблюдений в различные группы.

Модели, реализованные в модуле, являются линейными, а функции классификации и дискриминантные функции – линейными комбинациями наблюдаемых величин.

Выберем команду Discriminant Analysis через меню Statistics – Multivariate Exploratory Techniques (многомерные исследовательские методы)**.** В стартовой панели модуля выберем переменные для анализа.

В качестве *Grouping variable*(группирующие переменные) выберем переменную Groups(классы: отказ; высокий риск; средний риск; низкий риск; кредит). Группирующая переменная не должна входить в список независимых переменных. В качестве *Independent variable list*(список независимых переменных) выберем переменные – показатели с L1 по A6. Далее надо задать коды для значений группирующей переменной – кнопка Codes for grouping variables, выберем все 5 классов**.** В поле Method**(**метод**)** укажем метод дискриминантного анализа: ***Standard*** (стандартный). При этом методе все выбранные переменные будут одновременно включены в модель (уравнение). Рассмотрим окно результатов (рис.1)

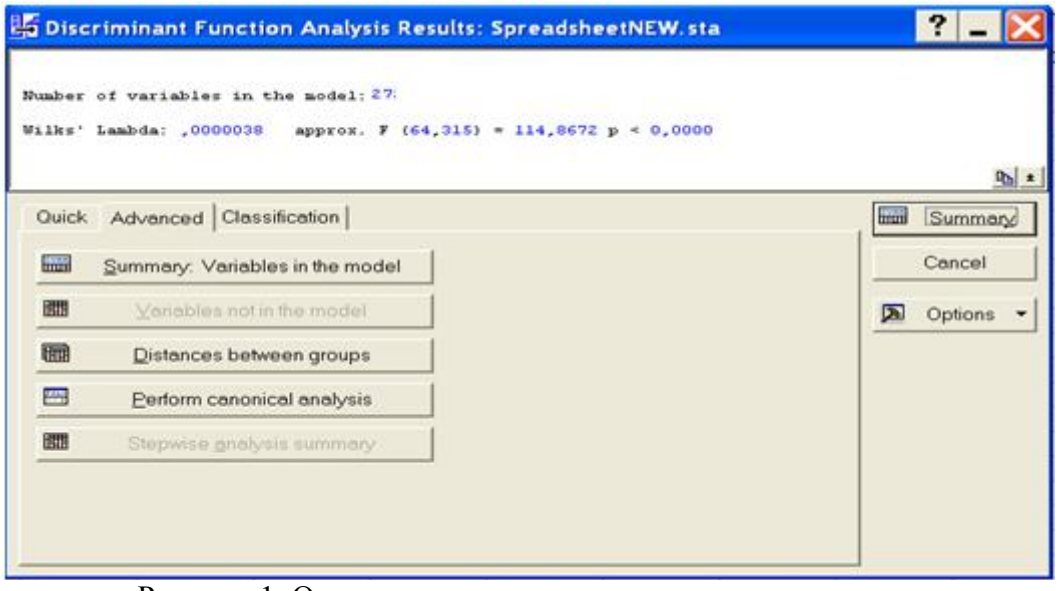

Рисунок 1. Окно результатов дискриминантного анализа

Информационная часть окна сообщает, что:

- *Number of variables in model* (число переменных в модели) равно 27;
- *Wilks 'Lambda* (значение лямбда Уилкса) равно 0,0000038;
- *approx. F(32,164)* (приближенное значение *F-статистики* с числом степеней свободы 64 и 315) равно 114,8672;
- *p* (уровень значимости *F-критерия)* меньше 0,0000.

Статистика лямбда Уилкса *(λ)* вычисляется как отношение детерминанта матрицы внутригрупповых дисперсий/ковариаций к детерминанту общей ковариационной матрицы. Значения *λ* принадлежат интервалу [0,1]. Значения λ равно 0,0000038, это говорит о хорошей дискриминации. В силу того факта, что значения λ лежащие около 0, свидетельствуют о хорошей дискриминации, а значения λ, лежащие около 1, свидетельствуют о плохой дискриминации.

Исследуем итоговую таблицу анализа данных, для этого выберем опцию Summary: Variables in the model (итоги: переменные, включенные в модель). Появится итоговая таблица анализа данных (рис.2):

	Discriminant Function Analysis Summary (Spreadsheet15) No. of vars in model: 16; Grouping: Grups (5 grps) Wilks' Lambda: ,00000 approx. F (64,315)=114,87 p<0,0000					
N=100	**Wilks' Lambda**	Partial Lambda	F-remove (4,80)	p-level	Toler.	1-Toler. (R-Sqr.)
L1	0,000005720	0,657565	10,41527	0,000001	0,729557	0,270443
L3	0,000005287	0,711360	8,11518	0,000015	0,714467	0,285533
P1	0,000003964	0,948750	1,08036	0,371856	0,752661	0,247339
F1	0,000006565	0,572909	14,90958	0,000000	0,843917	0,156083
F2	0,000005478	0,686626	9,12794	0,000004	0,921810	0,078190
F3	0,000006272	0,599707	13,34963	0,000000	0,689492	0,310509
F4	0,000004179	0,899931	2,22392	0,073720	0,834096	0,165904
R1	0,000006194	0,607232	12,93636	0,000000	0,885651	0,114349
R2	0,000004585	0,820369	4,37928	0,002974	0,894826	0,105174
R3	0,000003900	0,964404	0,73819	0,568680	0,643952	0,356048
R4	0,000004564	0,824113	4,26852	0,003503	0,730625	0,269375
R5	0,000004355	0,863638	3,15785	0,018329	0,708686	0,291314
A2	0,000004121	0,912726	1,91239	0,116400	0,783474	0,216526
A4	0,000005255	0,715664	7,94608	0,000019	0,679286	0,320714
A5	0,000006535	0,575519	14,75124	0,000000	0,692875	0,307125
A6	0,000005882	0,639386	11,28002	0,000000	0,612207	0,387793

Рисунок 2. Часть итоговой таблицы анализа данных

В первом столбце таблицы приведены значения *Wilks Lambda,* являющиеся результатом исключения соответствующей переменной из модели. Чем больше значение λ, тем более желательно присутствие этой переменной в процедуре дискриминации. Если сравнить порядок λ при делении множества объектов (предприятий) на 3 группы и на 5 групп, то порядок λ

Значение *Partial Lambda* (частная лямбда) есть отношение лямбда Уилкса после добавления соответствующей переменной к лямбде Уилкса до добавления этой переменной. Частная лямбда характеризует единичный вклад соответствующей переменной в разделительную силу модели. Чем меньше статистика лямбда Уилкса, тем больше вклад в общую дискриминацию, из таблицы видно, что переменная R3 дает вклад больше всех, затем переменные P1; A2; F4; R5; R4; R2 следующие по значимости – A4; L3; F2; L1; A6; R1; F3; а переменные A5 и F1 вносит в общую дискриминацию вклад меньше всех. Поэтому на этой стадии исследования можно заключить, что показатели рентабельности собственного капитала (R3), а также текущего коэффициента ликвидности (P1) и оборачиваемости активов (A2); являются главными переменными, которые позволяют производить дискриминацию между различными группами предприятий.

Толерантность *(Toler)* определяется как $1\text{-}R^2$, где R^2 – это коэффициент множественной корреляции данной переменной со всеми другими переменными в модели. Толерантность является мерой избыточности переменной в модели. Как видно из таблицы все переменные с большими значениями толерантности от 0,07 до 0,39, что намного больше нижней границы толерантности (0,01), значит все переменные успешно включены в модель. Так как переменные с толерантностью меньше заданного значения в модель не включаются. Чем меньше значение толерантности, тем избыточнее переменная в модели,

так как переменная несет малую дополнительную информацию. Меньше всех значение толерантности у показателей A6 и R3.

Для получения дальнейших результатов о природе дискриминации проведем ***канонический анализ***. Чтобы увидеть, как двадцать семь переменных разделяют различные совокупности (группы предприятий), надо вычислить дискриминантную функцию. Каждая последующая дискриминантная функция будет вносить все меньший и меньший вклад в общую дискриминацию. Максимальное число оцениваемых функций равно числу переменных (27) или числу совокупностей (5) минус один, в зависимости от того, какое число меньше. В нашем случае оцениваются четыре дискриминантные функции.

Выберем опцию ***Perform canonical analysis*** (выполнение канонического анализа), программа вычислит независимые (ортогональные) дискриминантные функции. Перейдем ко вкладке ***Quick***. Рассмотрим таблицу результатов с пошаговым критерием для канонических корней — дискриминантных функций. Для этого выберем кнопку *Summary: chi square tests of successive roots* (рис.3).

	Chi-Square Tests with Successive Roots Removed (Показатели)					
Roots Removed	**Eigen-value**	Canonicl R	Wilks' Lambda	Chi-Sqr.	df	p-level
0	1170,474	0,999573	0,000004	1105,435	64	0,000000
1	25,521	0,980966	0,004406	480,092	45	0,000000
2	5,305	0,917283	0,116854	189,994	28	0,000000
3	0,357	0,513009	0,736822	27,029	13	0,012329

Рисунок 3. Таблица результатов с пошаговым критерием для канонических корней

Первая строка дает критерий значимости для всех корней. Вторая строка содержит значимость корней, оставшихся после удаления первого корня и т.д. Таким образом, таблица позволит оценить, сколько значимых корней нужно интерпретировать. Все четыре дискриминантные функции статистически значимы.

Выберем опцию *Coefficients for canonical variables* (коэффициенты канонических переменных), появятся две таблицы с коэффициентами дискриминантных (канонических) функций. В первой таблице даны исходные (нестандартизованные) коэффициенты дискриминантных функций (рис. 4). Во второй таблице приведены стандартизованные коэффициенты дискриминантных функций. Эти коэффициенты, основанные на стандартизованных переменных, принадлежат к одной и то же шкале измерений (абсолютной), поэтому их можно сравнивать, чтобы определить величины и направления вкладов переменных в каждую каноническую функцию. Из таблицы видно, что наибольший вклад:

- в дискриминантную функцию 1 вносят переменные: F1; F4; R2,
- в дискриминантную функцию 2 – A6; A5; R4
- в дискриминантную функцию 3 – F3; A4; A2
- в дискриминантную функцию 4 – L3; A5; L1

В таблицах приведены собственные значения *(Eigenval)* для каждой дискриминантной функции и кумулятивная доля объясненной дисперсии *(Cum.Prop.),* накопленной каждой функцией.

Как видно:

- ✓ функция 1 ответственна за 97,40% объясненной дисперсии, т.е. 97,40% всей дискриминирующей мощности определяется этой функцией. Поэтому эта функция наиболее «важна»;
- ✓ функция 2 ответственна за 2,13% объясненной дисперсии;
- ✓ функция 3 ответственна за 0,44% объясненной дисперсии;
- ✓ функция 4 ответственна за 0.03% объясненной дисперсии.

Variable	Standardized Coefficients for Canonical Variables			
	Root 1	Root 2	Root 3	Root 4
L1	-0,453	-0,21427	-0,477260	0,331328
L3	-0,355	-0,12336	-0,498993	0,451553
P1	-0,151	-0,20533	0,070747	0,047624
F1	0,415	0,35220	-0,490107	-0,221436
F2	-0,364	-0,45188	-0,097907	0,102627
F3	-0,456	-0,02526	0,566972	-0,620889
F4	0,200	0,21852	-0,200443	0,034853
R1	-0,317	-0,08555	-0,620239	-0,214710
R2	-0,025	0,19004	-0,309051	-0,568357
R3	-0,095	-0,01709	0,208756	-0,188553
R4	-0,234	0,42774	-0,085202	-0,123335
R5	-0,208	0,32461	-0,187166	0,264662
A2	-0,212	0,00007	0,277897	-0,078467
A4	-0,353	0,41638	0,385544	-0,085626
A5	-0,468	0,55203	0,259480	0,409503
A6	-0,464	0,58486	0,196937	0,212580
Eigenval	1170,474	25,52108	5,305478	0,357180
Cum.Prop	0,974	0,99529	0,999703	1,000000

Рисунок 4. Таблица стандартизованных коэффициентов дискриминантных функций.

Выберем опцию ***Means of canonical variables*** (средние канонических переменных). Программа выведет таблицу (рис. 5) со средними значениями для дискриминантных функций, которые позволяют определить группы, лучше всего идентифицируемые конкретной дискриминантной функцией. Из таблицы видно, что

- дискриминантная функция 1 идентифицирует в основном группы 4 и 5, т.е. «ОТКАЗ» и «КРЕДИТ» (значение среднего значительно отличается от среднего РИСК),
- дискриминантная функция 2 — группы 2 и 5, т.е. «НИЗКИЙ РИСК» и «КРЕДИТ»
- дискриминантная функция 3 — группы 1 и 2, т.е. «НИЗКИЙ РИСК» и «СРЕДНИЙ РИСК»
- дискриминантная функция 4 — группы 1 и 3, т.е. «СРЕДНИЙ РИСК» и «ВЫСОКИЙ РИСК».

Group	Means of Canonical Variables			
	Root 1	Root 2	Root 3	Root 4
G_1:1	4,6614	-4,40627	3,04403	-0,674617
G_2:2	-20,3352	-6,57205	-2,99467	0,151156
G_3:3	27,4663	0,22240	1,47813	0,989590
G_4:4	40,4749	4,77637	-2,10824	-0,488722
G_5:5	-52,2675	5,97955	0,58076	0,022593

Рисунок 5. Таблица средних канонических переменных

Судить о результатах разделения программой наблюдений по группам, удобней по диаграмме рассеяния (рис.6а). Выберем опцию ***Scatterplot of canonical scores*** (диаграмма рассеяния для канонических значений).

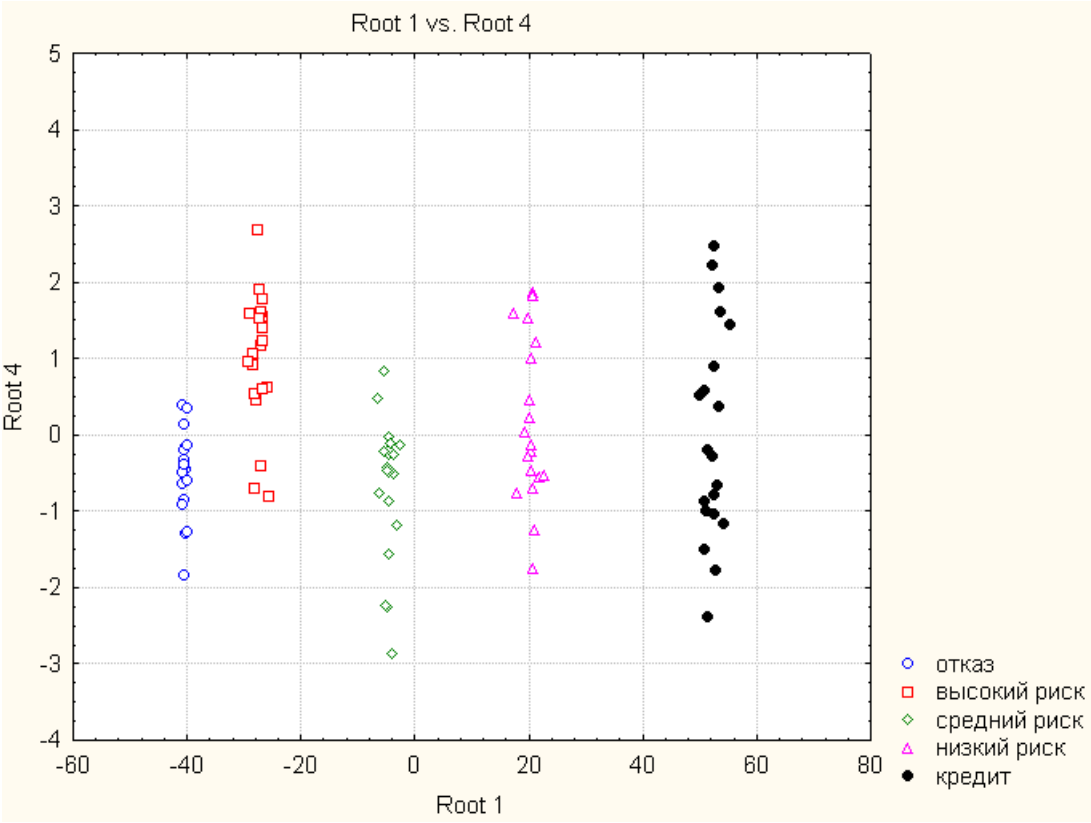

Рисунок 6а. Диаграмма рассеяния канонических значений для пар значений дискриминантных функций 1 и 4.

На диаграмме видно, что наблюдения (предприятия), принадлежащие одинаковым группам, локализованы в определенных областях плоскости, при этом расстояние между центроидами групп Отказ и Высокий риск; Высокий риск и Средний риск; Средний риск и Низкий риск; Низкий риск и Кредит почти одинаковые это может говорить о том, что «соседние» группы одинаково попарно отличны друг от друга. Самое большее расстояние между группами Отказ и Кредит, они наиболее различны между собой.

Из диаграммы видно, что группы ОТКАЗ и КРЕДИТ наиболее отдалены друг от друга, т.е. показателям в этих группах соответствуют большие значения корня 1(Root 1). Поэтому дискриминантная функция 1 главным образом дискриминирует показатели между этими группами и группами РИСКА. Дискриминантная функция 4, по-видимому, дает основную дискриминацию между показателями групп РИСКА и другими группами. Однако дискриминация не так отчетлива, как для дискриминантной функции 1. Из диаграммы видно, что дискриминация по дискриминантной функции 1 более сильная, чем по дискриминантной функции 4.

Аналогичным образом производя анализы остальных дискриминантных функций, можем сделать вывод, что наиболее сильной из всех является дискриминантная функция 1, а наиболее слабой - дискриминантная функции 4. Для этого достаточно сравнить диаграммы рассеяния канонических значений для пар значений дискриминантных функций 1 и 2; 1 и 3, а так же диаграммы рассеяния канонических значений для пар значений дискриминантных функций 2 и 4; 3 и 4. (рис. 6б-6д.):

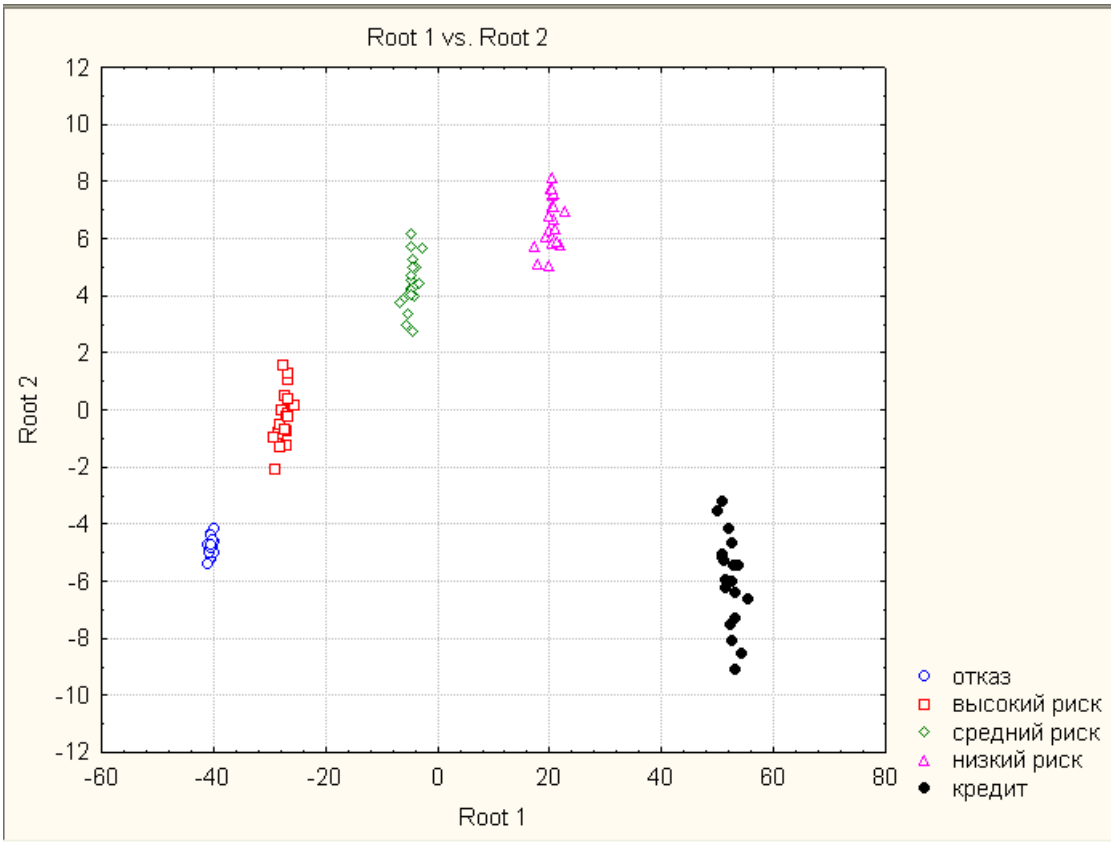

Рисунок 6б. Диаграмма рассеяния канонических значений для пар значений дискриминантных функций 1 и 2.

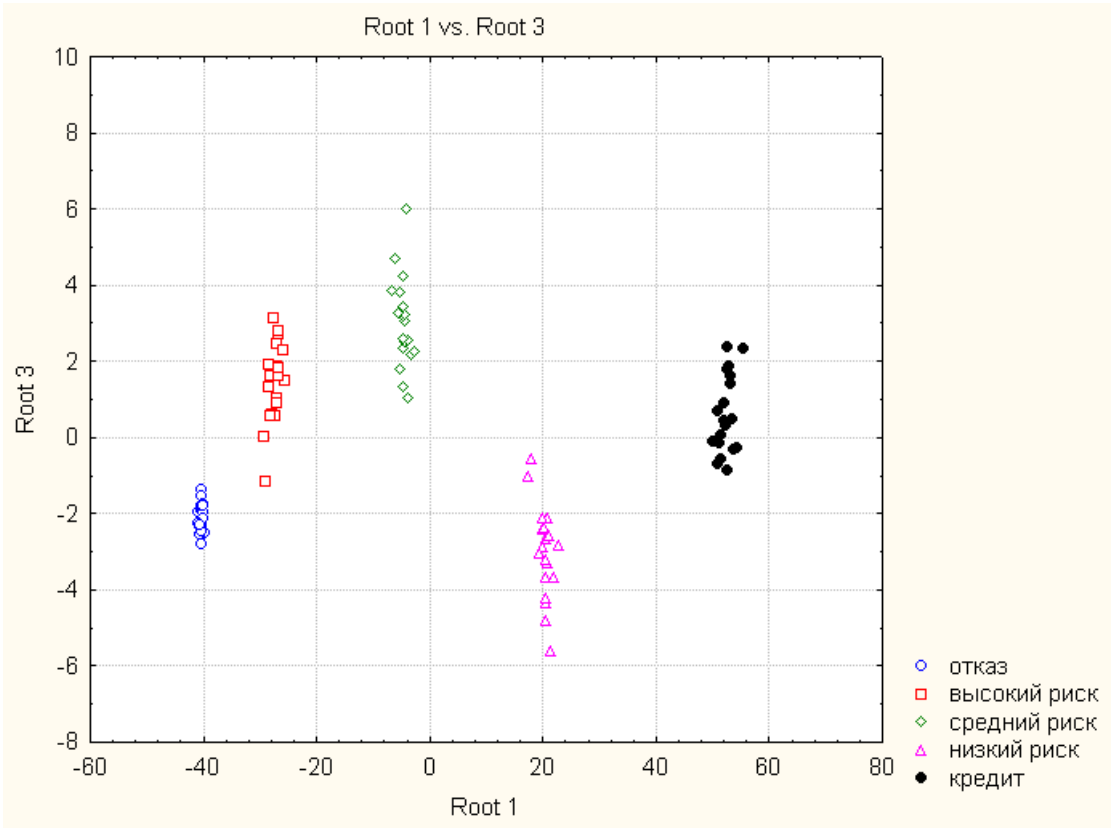

Рисунок 6в. Диаграмма рассеяния канонических значений для пар значений дискриминантных функций 1 и 3.

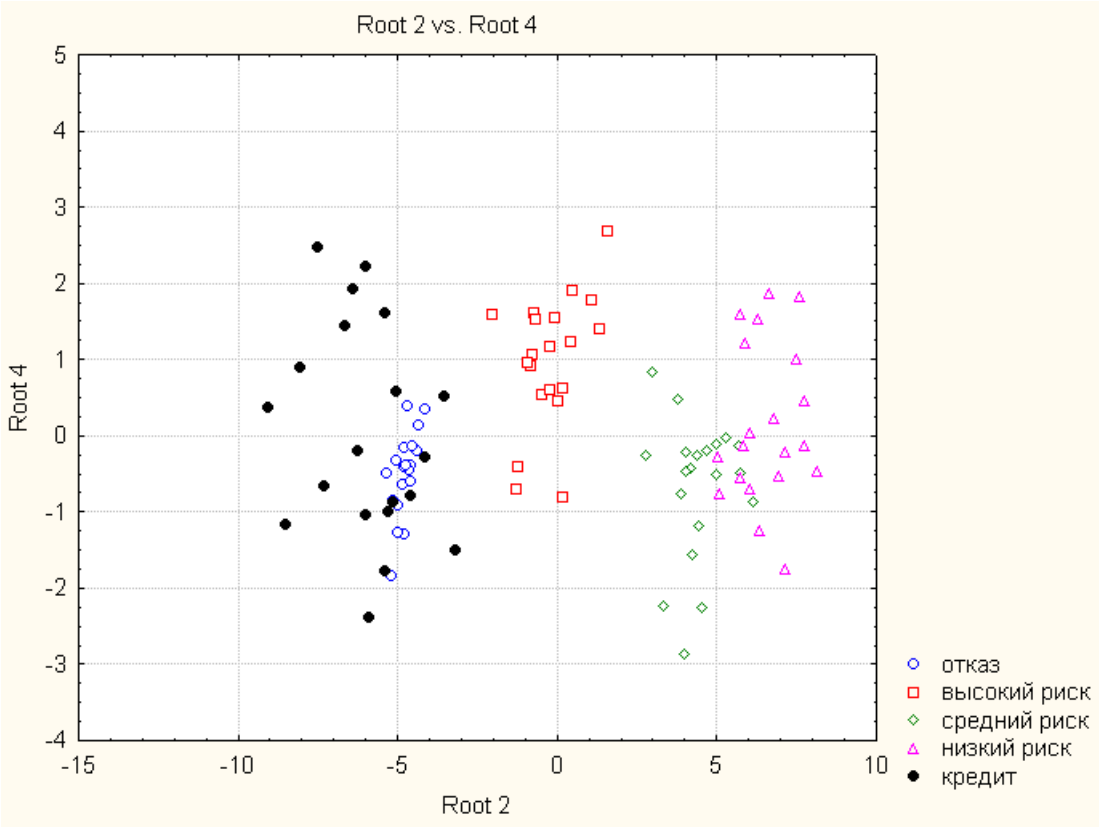

Рисунок 6г. Диаграмма рассеяния канонических значений для пар значений дискриминантных функций 2 и 4.

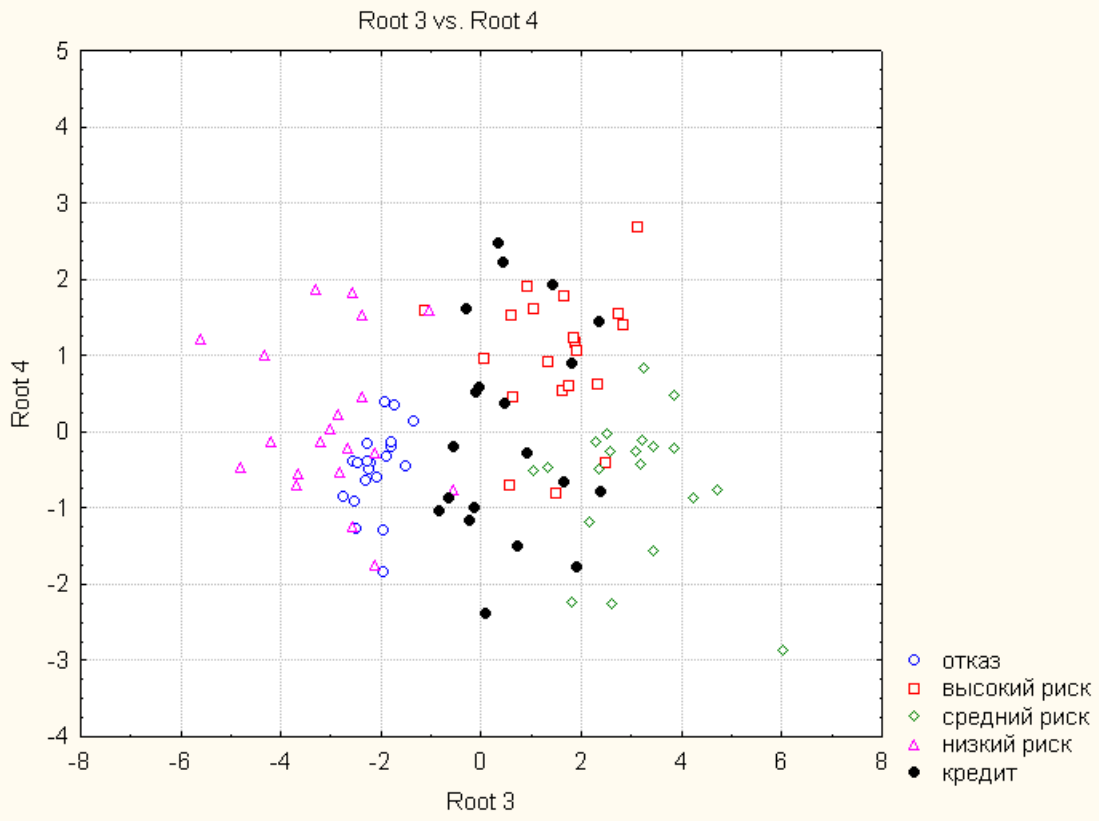

Рисунок 6д. Диаграмма рассеяния канонических значений для пар значений дискриминантных функций 3 и 4.

Выделим интервалы по значениям дискриминантной функции 1 для групп:

- значения принадлежащие группе «КРЕДИТ» будут находятся в [32; +∞),
- значения принадлежащие группе «НИЗКИЙ РИСК» будут находятся в [12; 32),
- значения принадлежащие группе «СРЕДНИЙ РИСК» будут находятся в (-16; 12),
- значения принадлежащие группе «ВЫСОКИЙ РИСК» будут находятся в (-35; -16],
- соответственно значения принадлежащие группе «ОТКАЗ» будут из (-∞; -35].

Вернемся в окно результатов ***Discriminant Function Analysis Results*** и активизируем вкладку ***Classification*** (классификация). Выберем опцию ***Classification functions***(функции классификации). Функции классификации – это линейные функции, которые вычисляются для каждой группы и могут быть использованы для классификации наблюдений (предприятий). Предприятие приписывают той группе, для которой классификационная функция имеет наибольшее значение. В таблице (рис.7), приведены коэффициенты и свободные члены при переменных линейных функций.

	Classification Functions; grouping: Grups (Spreadsheet15)				
Variable	**отказ p=,20000**	высокий риск p=,20000	средний риск p=,20000	низкий риск p=,20000	кредит p=,20000
L1	-23,96	42,25	159,78	335,03	452,55
L3	0,78	0,97	1,30	1,88	2,25
P1	97,48	126,34	165,95	200,23	222,49
F1	490,24	379,11	238,59	136,43	7,44
F2	458,56	663,69	976,95	1311,53	1485,61
F3	-91,62	-51,97	16,86	58,48	148,99
F4	837,56	752,88	634,85	550,82	464,20
R1	10,12	11,72	17,41	26,89	32,46
R2	6,65	1,46	1,76	5,10	9,56
R3	-25,56	-23,58	-20,38	-19,28	-15,31
R4	0,92	1,23	3,02	5,81	12,38
R5	-5,08	-4,02	-1,22	5,31	16,50
A2	190,60	341,86	566,34	714,06	1037,24
A4	-39,60	-26,75	-4,42	13,55	71,68
A5	-18,43	1,18	31,28	66,87	154,21
A6	20,57	27,73	40,49	55,67	92,87
Constant	-1533,29	-1262,78	-1358,78	-2297,38	-4236,86

Рисунок 7. Таблица с коэффициентами функций классификации

Функции классификации для групп «Отказ», «Высокий риск», «Средний риск», «Низкий риск», «Кредит» имеют вид:

Z1(«Отказ») = -23,96 L1 + 0,78L3 + 97,48 P1 + 490,24 F1 + 458,56F2 – 91,62 F3 + 837,56F4 + 10,12R1 + 6,65R2 – 25,56R3 + 0,92R4 – 5,08R5 + 190,60A2 – 39,60A4 – 18,43A5 + 20,57A6 –1533,29

Z2(«Высокий риск») = 42,25L1 +0,97 L3 + 126,34P1 + 379,11F1 + 663,69F2 − 51,97F3 + 752,88F4 + 11,72R1 + 1,46R2 − 23,58R3 + 1,23R4 – 4,02R5 + 341,86A2 – 26,75A4 + 1,18A5 + 27,73A6 – 1262,78

Z3(«Средрий риск») = 159,78L1 + 1,3L3 + 165,95P1 + 238,59F1 + 976,95F2 + 16,86F3 + 634,85F4 + 17,41R1 + 1,76R2 − 20,38R3 + 3,02R4 – 1,22R5 + 566,34A2 – 4,42A4 + 31,28A5 + 40,49A6 – 1358,78

Z4(«Низкий риск») = 335,03L1 +1,88 L3 + 200,23P1 + 136,43F1 + 1311,53F2 + 58,48F3 + 550,82F4 + 26,89R1 + 5,10R2 − 19,28R3 + 5,81R4 + 5,31R5 + 714,06A2 + 13,55A4 + 66,87A5 + 55,67A6 – 2297,38

Z5(«Кредит») = 452,55L1 + 2,25L3 + 222,49P1 + 7,44F1 + 1485,61F2 + 148,99F3 + 464,20F4 + 32,46R1 + 9,56R2 − 15,31R3 + 12,38R4 + 16,50R5 + 1037,24A2 + 71,68A4 + 154,21A5 + 92,87A6 – 4236,86

Выберем опцию ***Classification matrix*** (матрица классификации). Матрица (рис. 8) содержит информацию о количестве и проценте корректно классифицированных наблюдений в каждой группе. Строки матрицы — исходные классы, столбцы – предсказанные классы. Как видно из таблицы, исходные и предсказанные классы полностью совпали.

Classification Matrix (Spreadsheet15)
Rows: Observed classifications
Columns: Predicted classifications

Group	Percent Correct	отказ p=,20000	высокий риск p=,20000	средний риск p=,20000	низкий риск p=,20000	кредит p=,20000
отказ	100,0000	20	0	0	0	0
высокий риск	100,0000	0	20	0	0	0
средний риск	100,0000	0	0	20	0	0
низкий риск	100,0000	0	0	0	20	0
кредит	100,0000	0	0	0	0	20
Total	100,0000	20	20	20	20	20

Рисунок 8. Матрица классификации

Выберем опцию ***Squared Mahalanobis distances*** (квадраты расстояний Махаланобиса). Будет выведена таблица квадратов расстояний Махалонобиса каждого наблюдения от центра до группы (рис. 9).

Squared Mahalanobis Distances from Group Centroids (Spreadsheet15)
Incorrect classifications are marked with *

Case	Observed Classif.	отказ p=,20000	высокий риск p=,20000	средний риск p=,20000	низкий риск p=,20000	кредит p=,20000
C_10	отказ	1,430	222,398	1429,824	3879,320	8682,480
C_11	отказ	1,311	208,832	1399,352	3829,308	8621,475
C_12	отказ	1,602	201,396	1375,163	3799,838	8563,522
C_13	отказ	1,886	226,016	1442,723	3902,389	8702,324
C_14	отказ	2,350	196,691	1381,897	3823,524	8613,205
C_15	отказ	2,356	198,781	1379,096	3814,324	8587,700
C_16	отказ	0,872	204,034	1390,207	3817,159	8593,262
C_17	отказ	0,714	192,456	1358,194	3769,409	8526,757
C_18	отказ	0,916	207,383	1400,164	3840,210	8616,792
C_19	отказ	1,392	211,299	1408,373	3850,350	8645,616
C_20	отказ	2,505	193,865	1366,028	3786,745	8551,722
C_21	высокий риск	216,124	5,653	540,505	2319,352	6321,970
C_22	высокий риск	150,533	21,854	670,564	2533,128	6650,947
C_23	высокий риск	180,555	2,757	598,036	2459,376	6544,543
C_24	высокий риск	176,178	4,903	604,600	2462,334	6554,537
C_25	высокий риск	240,460	6,789	512,970	2281,575	6287,203
C_26	высокий риск	239,629	6,597	521,692	2309,290	6304,384
C_27	высокий риск	264,725	13,624	486,923	2228,654	6180,479
C_28	высокий риск	191,849	6,934	577,534	2399,534	6486,327
C_29	высокий риск	223,936	8,027	540,944	2336,080	6354,492
C_30	высокий риск	220,255	10,218	538,966	2341,644	6322,211

Рисунок 9. Фрагмент таблицы квадратов расстояний Махалонобиса

Эти расстояния аналогичны квадратам евклидовых расстояний, но учитывают корреляции между переменными в модели. Наблюдение приписывают группе, к которой оно ближе всего. Как видно первые 20 наблюдений ближе всего к группе «Отказ». Наблюдения, которые не удалось правильно классифицировать, также помечены *. Таких наблюдений нет.

Перейдем на вкладку ***Canonical scores*** (канонические значения), и выберем опцию ***Canonical scores for each case*** (канонические значения для каждого наблюдения). Появится таблица со значениями дискриминантных функций для каждого наблюдения. Сохраним эти значения через ***Save canonical scores*** (сохранить канонические значения) и используем механизм двойной записи, рассмотренный раннее, для переменной Groops. **(**рис. 10**)**

Data: SpreadsheetNEW.sta* (19v by 100c)

Spreadsheet1 in MYbase.stw

	8 R1	9 R2	10 R3	11 R4	12 R5	13 A2	14 A4	15 A5	16 A6	17 CASE_NO	18 CLUSTER	19 DISTANCE
C_10	-1,135	-0,838	-1,003	-1,022	-1,056	-1,166	-0,912	-0,941	-0,887	10	отказ	0,04
C_11	-1,140	-0,813	-1,034	-1,033	-1,058	-1,162	-0,899	-0,984	-0,881	11	отказ	0,04
C_12	-1,128	-0,766	-1,048	-1,023	-1,048	-1,136	-0,878	-0,963	-0,865	12	отказ	0,04
C_13	-1,130	-0,949	-1,030	-1,019	-1,060	-1,129	-0,881	-0,943	-0,898	13	отказ	0,04
C_14	-1,119	-1,091	-1,041	-1,017	-1,032	-1,176	-0,898	-0,963	-0,899	14	отказ	0,05
C_15	-1,099	-0,984	-1,048	-1,038	-1,038	-1,225	-0,886	-0,959	-0,882	15	отказ	0,04
C_16	-1,145	-0,874	-1,011	-1,013	-1,045	-1,152	-0,871	-0,930	-0,878	16	отказ	0,03
C_17	-1,119	-0,839	-1,020	-1,026	-1,043	-1,163	-0,885	-0,978	-0,865	17	отказ	0,03
C_18	-1,124	-0,916	-1,014	-1,024	-1,034	-1,155	-0,885	-0,932	-0,872	18	отказ	0,03
C_19	-1,133	-0,900	-1,014	-1,028	-1,051	-1,189	-0,898	-0,975	-0,870	19	отказ	0,03
C_20	-1,117	-1,061	-1,051	-1,025	-1,060	-1,089	-0,915	-0,940	-0,899	20	отказ	0,05
C_21	-0,892	-0,996	-0,775	-0,834	-0,722	-0,697	-0,706	-0,672	-0,654	21	высокий риск	0,06
C_22	-0,902	-0,905	-0,757	-0,726	-0,695	-0,884	-0,713	-0,750	-0,630	22	высокий риск	0,14
C_23	-1,009	-0,911	-0,769	-0,851	-0,848	-0,740	-0,690	-0,696	-0,649	23	высокий риск	0,05
C_24	-0,924	-0,992	-0,784	-0,839	-0,911	-0,676	-0,704	-0,742	-0,580	24	высокий риск	0,06
C_25	-0,893	-1,015	-0,819	-0,794	-0,778	-0,827	-0,686	-0,672	-0,771	25	высокий риск	0,08
C_26	-0,965	-1,002	-0,861	-0,798	-1,018	-0,738	-0,734	-0,621	-0,579	26	высокий риск	0,08
C_27	-0,959	-0,850	-0,776	-0,880	-0,841	-0,786	-0,677	-0,692	-0,621	27	высокий риск	0,09
C_28	-0,921	-0,775	-0,791	-0,823	-0,939	-0,832	-0,749	-0,751	-0,701	28	высокий риск	0,07
C_29	-0,802	-1,077	-0,874	-0,766	-0,831	-0,778	-0,662	-0,793	-0,651	29	высокий риск	0,09
C_30	-0,967	-0,775	-0,681	-0,849	-0,710	-0,751	-0,640	-0,734	-0,692	30	высокий риск	0,10

Рисунок 10. Сохраненные канонические значения для каждого наблюдения

В заключении дискриминантного анализа была получина таблица значений апостериорных вероятностей, то есть вероятностей принадлежности каждого предприятия к пяти группам риска банкротства (таблица 11). Клиент приписывается к той группе риска, для которой имеется наибольшая апостериорная вероятность классификации.

Posterior Probabilities (NEW BASE_1000.sta)
Incorrect classifications are marked with *

Case	Observed Classif.	отказ p=.20000	Высокий риск p=.20000	Средний риск p=.20000	Низкий риск p=.20000	Кредит p=.20000
183	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
184	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
185	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
186	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
187	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
188	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
189	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
190	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
191	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
192	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
193	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
194	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
195	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
196	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
197	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
198	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
199	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
200	отказ	1.000000	0.000000	0.000000	0.000000	0.000000
201	Высокий риск	0.000000	1.000000	0.000000	0.000000	0.000000
202	Высокий риск	0.000000	1.000000	0.000000	0.000000	0.000000
203	Высокий риск	0.000000	1.000000	0.000000	0.000000	0.000000
204	Высокий риск	0.000000	1.000000	0.000000	0.000000	0.000000
205	Высокий риск	0.000000	1.000000	0.000000	0.000000	0.000000
206	Высокий риск	0.000000	1.000000	0.000000	0.000000	0.000000
207	Высокий риск	0.000000	1.000000	0.000000	0.000000	0.000000
208	Высокий риск	0.000000	1.000000	0.000000	0.000000	0.000000

Таблица 11. Значения апостериорных вероятностей для предприятий модельной базы.

На данном этапе анализа была произведена классификация предприятий тестовой выборки. Исходная модельная база клиентов была расширена тестовой выборкой. Таблица с апостериорными вероятностями позволила определить уровень риска банкротства каждого из вновь введенного предприятия.

Сумма вероятностей для каждого предприятия равна 1. Чем дальше наблюдение расположено от центра группы, тем менее вероятно, что оно принадлежит к этой группе. Наблюдение приписывают к той группе, для которой имеется наибольшая апостериорная вероятность классификации.

Априорные вероятности могут быть заданы пользователем, могут быть равны для всех групп, могут быть пропорциональны размерам групп. Из таблицы видно, что положение предприятий в группах в высшей степени устойчиво, так как апостериорные вероятности близки либо к 1, либо к 0.

Программа в процессе *ДА*, «обучившись» по таблице исходных данных, может прогнозировать уровень риска банкротства для любого нового предприятия. Для этого надо в конец таблицы исходных данных добавить новые наблюдения, например, при помощи контекстного меню (правая кнопка мыши) и команды *Добавить наблюдение*. При этом предполагая, что неизвестен уровень риска банкротства предприятия, не указываем группу. Запустив процедуру *ДА* и воспользовавшись кнопкой *Апостериорные вероятности*, мы получим вероятности принадлежности предприятия к той или иной группе уровня риска банкротства.