

РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ

Факультет физико-математических и естественных наук

Кафедра прикладной информатики и теории вероятностей

Практикум по математической статистике

Лабораторная работа №5

Тема: «Дискриминантный анализ»

Вариант 10

Выполнил

Студент: Феокистов Владислав

Группа: НПМбд-01-196

№ с/б: 1032192939

Преподаватель: Матюшенко Сергей Иванович

МОСКВА

2022 г.

Цель работы: приобрести практические навыки применения дискриминантного анализа для решения конкретных задач с использованием статистического пакета SPSS.

Ход работы:

1. Изучил теоретические основы дискриминантного анализа, используя материалы учебного пособия.
2. Разобрал пример использования SPSS для реализации дискриминантного анализа.
3. Запустил программу SPSS и введем исходные данные по обучающей выборке.

	A	B	C	D	E	F	G	H
1	№	Брался ли кредит ранее ¹ (Да, Нет)	Среднемесячный доход семьи заемщика, тыс. руб.	Период погашения кредита, лет	Размер кредита, тыс. руб.	Состав семьи заемщика, чел.	Возраст заемщика, лет	Вероятность погашения кредита
2								
3								
4	1	2	3	4	5	6	7	8
5	Данные для построения дискриминантной функции (Обучающая выборка)							
6	id	X1	X2	X3	X4	X5	X6	X7
7	1	1	27,6	7	140	4	46	2
8	2	1	25,6	8	190	5	37	2
9	3	1	32,75	8	170	6	43	2
10	4	1	36,5	7	290	4	52	3
11	5	1	26,45	6	200	3	44	2
12	6	1	38,75	9	390	5	49	3
13	7	1	24,35	6	150	3	53	2
14	8	2	23,3	7	380	3	43	2
15	9	1	32,25	6	180	2	49	3
16	10	2	19,6	4	240	3	29	1
17	11	1	37,95	7	190	5	45	3
18	12	1	37,15	6	220	4	42	3
19	13	2	21,9	4	270	2	43	2
20	14	1	25,9	5	140	3	57	3
21	15	2	19,9	3	110	4	55	1
22	16	2	17,3	6	160	3	59	1
23	17	2	19,35	4	150	2	56	1
24	18	2	22,9	3	170	2	38	2
25	19	2	26,45	6	240	4	34	1

Открытие файлов Excel

C:\Users\phoen\Documents\GitHub\Math-statistics\Lab05\2.1.10. Данные для д.з. №5.xls

☒ Читать имена переменных из первой строки данных

Лист: Зад.4 Обуч. подмнож [A1:H36]

Диапазон: B6:H36

Максимальная ширина текстовых столбцов: 32767

OK Отмена Справка

	X1	X2	X3	X4	X5	X6	X7
1	1	27,6	7	140	4	46	2
2	1	25,6	8	190	5	37	2
3	1	32,8	8	170	6	43	2
4	1	36,5	7	290	4	52	3
5	1	26,5	6	200	3	44	2
6	1	38,8	9	390	5	49	3
7	1	24,4	6	150	3	53	2
8	2	23,3	7	380	3	43	2
9	1	32,3	6	180	2	49	3
10	2	19,6	4	240	3	29	1
11	1	38,0	7	190	5	45	3
12	1	37,2	6	220	4	42	3
13	2	21,9	4	270	2	43	2
14	1	25,9	5	140	3	57	3
15	2	19,9	3	110	4	55	1
16	2	17,3	6	160	3	59	1
17	2	19,4	4	150	2	56	1
18	2	22,9	3	170	2	38	2
19	2	26,5	6	240	4	34	1
20	1	35,3	8	180	5	46	3
21	2	20,4	4	180	2	46	1
22	2	28,8	2	140	2	58	2
23	2	29,8	9	500	2	37	3
24	2	18,8	11	420	5	45	1

	Имя	Тип	Ширина	Десятич.	Метка	Значения	Пропуски	Столбцы	Выравнивание	Шкала
1	X1	Числовая	11	0	История	{1, Да}...	Нет	11	По право...	Номинальная...
2	X2	Числовая	11	1	Доход	Нет	Нет	11	По право...	Количество...
3	X3	Числовая	11	0	Срок	Нет	Нет	11	По право...	Количество...
4	X4	Числовая	11	0	Кредит	Нет	Нет	11	По право...	Количество...
5	X5	Числовая	11	0	Семья	Нет	Нет	11	По право...	Количество...
6	X6	Числовая	11	0	Возраст	Нет	Нет	11	По право...	Количество...
7	X7	Числовая	11	0	Вероятность	{1, Низкая}...	Нет	11	По право...	Номинальная...

Далее сделаем дискриминантный анализ.

Дискриминантный анализ

Группировать по:
X7(1 3)

Задать диапазон...

Независимые:
История [X1]
Доход [X2]
Срок [X3]

☒ Принудительное включение
☐ Шаговый отбор

Переменная отбора наблюдений:
Значение...

Статистики...
Метод...
Классифицировать...
Сохранить...

OK Вставка Сброс Отмена Справка

Дискриминантный анализ: Статистики

Описательные
☒ Средние
☒ Однофакторный дисперсионный анализ
☐ M Бокса

Кoeffициенты функции
☐ Фишера
☒ Нестандартизованные

Матрицы
☒ Внутригрупповая корреляция
☐ Внутригрупповая ковариация
☐ Групповые ковариации
☐ Общая ковариация

Продолжить Отмена Справка

Дискриминантный анализ: Классификация

Априорные вероятности
☐ Все группы равны
☒ Вычислить по размерам групп

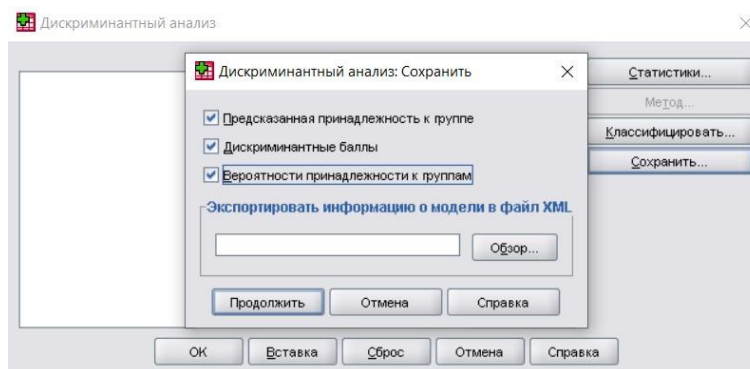
Ковариационная матрица
☒ Внутригрупповая
☐ Отдельно по группам

Вывести на дисплей
☒ Поточечные результаты
☐ Ограничиться первыми:
☒ Итоговая таблица
☒ Скользящий контроль

Графики
☒ Объединенные группы
☒ Для отдельных групп
☒ Территориальная карта

☐ Заменить пропущенные значения средним

Продолжить Отмена Справка



После нажатия на кнопку «ОК» получаем следующий вывод:

Сводка результатов обработки наблюдений				Групповые статистики					
Невзвешенные наблюдения		N	Процент	Вероятность	Среднее	Стд. отклонение	Кол-во валидных (искл. целиком)		
Валидные		30	100,0				Невзвешенные	Взвешенные	
Исключенные	Пропущенные или лежащие вне диапазона коды группирующей переменной	0	,0	Низкая	История	2,000	9	9,000	
					Доход	20,444	9	9,000	
					Срок	5,444	9	9,000	
					Кредит	218,889	9	9,000	
	По крайней мере одна пропущенная дискриминантная переменная	0	,0		Семья	3,111	9	9,000	
					Возраст	44,778	9	9,000	
	Оба групповых кода пропущены или лежат вне диапазона, и отсутствует по крайней мере одна дискриминантная переменная.	0	,0	Средняя	История	1,417	12	12,000	
					Доход	26,242	12	12,000	
					Срок	5,333	12	12,000	
					Кредит	211,667	12	12,000	
					Семья	3,750	12	12,000	
					Возраст	45,750	12	12,000	
	Итого искл.	0	,0	Высокая	История	1,111	9	9,000	
					Доход	34,089	9	9,000	
					Срок	7,222	9	9,000	
					Кредит	262,222	9	9,000	
					Семья	3,778	9	9,000	
					Возраст	47,889	9	9,000	
Всего набл.		30	100,0	Итого	История	1,500	30	30,000	
					Доход	26,857	30	30,000	
					Срок	5,933	30	30,000	
					Кредит	229,000	30	30,000	
					Семья	3,567	30	30,000	
					Возраст	46,100	30	30,000	

Из данных таблицы «Критерий равенства групповых средних» следует, что переменные «Кредит» («Размер кредита»), «Семья» («Состав семьи заемщика»), «Возраст» («Возраст заемщика») незначимо различаются по группам, поскольку для них уровень значимости $Z_{\text{нч.}} > 0.05$, поэтому классификацию заемщиков целесообразно проводить по первым двум переменным: «История» («Брался ли кредит») и «Доход» («Среднемесячный доход семье заемщика»).

Критерий равенства групповых средних					
	Лямбда Уилкса	F	ст.св1	ст.св2	Знч.
История	,507	13,106	2	27	,000
Доход	,291	32,831	2	27	,000
Срок	,838	2,619	2	27	,091
Кредит	,944	,796	2	27	,461
Семья	,948	,741	2	27	,486
Возраст	,978	,305	2	27	,740

Анализ матрицы коэффициентов в таблице «Объединенные внутригрупповые матрицы» свидетельствует об отсутствии мультиколлинеарности, поэтому коэффициенты корреляции малы.

Объединенные внутригрупповые матрицы

		История	Доход	Срок	Кредит	Семья	Возраст
Корреляция	История	1,000	-,455	-,141	,486	-,626	-,234
	Доход	-,455	1,000	,026	-,176	,621	,053
	Срок	-,141	,026	1,000	,583	,341	-,352
	Кредит	,486	-,176	,583	1,000	-,037	-,469
	Семья	-,626	,621	,341	-,037	1,000	,099
	Возраст	-,234	,053	-,352	-,469	,099	1,000

Данные таблицы «Собственные значения» показывают, что первая функция учитывает 95,2% дисперсии, а корреляция между исходными данными и данными, полученными по модели, высокая и составляет 0,929. Для второй функции эти значения намного меньше.

Собственные значения

Функция	Собственное значение	% объясненной дисперсии	Кумулятивный %	Каноническая корреляция
1	6,327 ^а	95,2	95,2	,929
2	,318 ^а	4,8	100,0	,491

а. В анализе использовались первые 2 канонические дискриминантные функции.

Оценка значимости дискриминантных функций проводится по коэффициенту Уилкса (λ). Из данных таблицы «Лямбда Уилкса» видно, что для первой функции значимость Знч. < 0,001, следовательно, она позволяет значимо и надежно дискриминировать наблюдения. В то же время значимость второй функции составляет лишь 0,239. Поэтому в дальнейшем для классификации целесообразно использовать только первую дискриминантную функцию.

Лямбда Уилкса

Проверка функции(й)	Лямбда Уилкса	Хи-квадрат	ст.св.	Знч.
от 1 до 2	,104	55,549	12	,000
2	,759	6,757	5	,239

Нормированные коэффициенты канонической дискриминантной функции

	Функция	
	1	2
История	-,963	1,235
Доход	1,039	,543
Срок	,019	1,276
Кредит	,807	-,837
Семья	-1,177	-,310
Возраст	,281	,383

Структурная матрица

	Функция	
	1	2
Доход	,620 [*]	-,012
Возраст	,059 [*]	,036
История	-,375	,507 [*]
Срок	,152	,388 [*]
Кредит	,080	,244 [*]
Семья	,076	-,243 [*]

Объединенные внутригрупповые корреляции между дискриминантными переменными и нормированными каноническими дискриминантными функциями. Переменные упорядочены по абсолютной величине корреляций внутри функции.

* Максимальная по абсолютной величине корреляция между переменными и дискриминантными функциями.

Формально по данным таблицы «Коэффициенты канонической дискриминантной функции» можно построить две дискриминантные функции:

$$D_1(X) = -4,286 - 2,566x_1 + 0,290x_2 + 0,009x_3 + 0,008x_4 - 0,876x_5 + 0,032x_6;$$

$$D_2(X) = -11,943 + 3,291x_1 + 0,151x_2 + 0,631x_3 - 0,009x_4 - 0,231x_5 + 0,044x_6$$

Коэффициенты канонической дискриминантной функции

	Функция	
	1	2
История	-2,566	3,291
Доход	,290	,151
Срок	,009	,631
Кредит	,008	-,009
Семья	-,876	-,231
Возраст	,032	,044
(Константа)	-4,286	-11,943

Ненормированные
коэффициенты

Однако поскольку значимость второй функции более 0,001, ее для дискриминации использовать нецелесообразно.

Координаты центроидов по группам приведены в таблице «Функции в центроидах групп». Они используются для нанесения центроидов на карту восприятия.

Функции в центроидах групп

Вероятность	Функция	
	1	2
Низкая	-2,873	,503
Средняя	-,289	-,652
Высокая	3,258	,366

Ненормированные
канонические
дискриминантные функции
вычислены в центроидах групп.

Зная координаты центроидов по группам и формулу дискриминантной функции $D_1(X)$, можно предсказать вероятность погашения кредита.

H4														
= -4,286 - 2,566 * B4 + 0,29 * C4 + 0,009 * D4 + 0,008 * E4 - 0,876 * F4 + 0,032 * G4														
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	№	Брался ли	Средне-	Период	Размер кредита,	Состав	Возраст							
	вариант	кредит ранее?	месячный доход	погашения	тыс. руб.	семьи	займщик	D1(X)	Расстояние до 1	Расстояние до 2	Расстояние до 3	Предсказанная группа		
	a	(Да -1, Нет-2)	семьи	кредита,		займщик	а, лет							D1(X)
			займщика,	лет		а, чел.								
			тыс. руб.											
1	10	2	19,99	5	270	3	28	-3,15237	0,279367595	2,863367595	6,410367595	1		1 -2,873
2	11	1	23,82	7	350	5	45	-0,0229	2,850099704	0,286099704	3,280900296	2		2 -0,289
3	12	1	16,64	3	320	4	42	-1,60144	1,271563097	1,312436903	4,859436903	1		3 3,258

Классификационные статистики

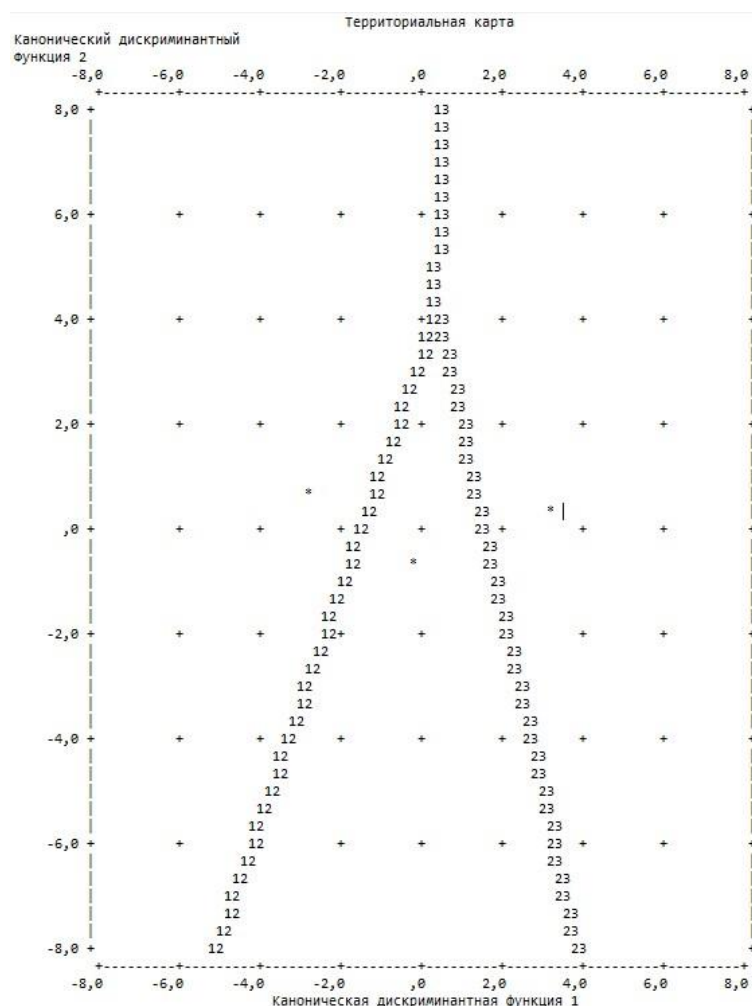
Сводка классификации	
Обработано	30
Исключенные2	0
Пропущенные или лежащие вне диапазона коды групп2	0
По крайней мере одна дискриминантная переменная пропущена.	
Используется в выводе	30

Априорные вероятности для групп

Вероятность	Априорные	Наблюдения, использованные в анализе	
		Невзвешенные	Взвешенные
Низкая	,300	9	9,000
Средняя	,400	12	12,000
Высокая	,300	9	9,000
Итого	1,000	30	30,000

Карта восприятий визуализирует разделение наблюдений функциями. Так, первая функция $D_1(X)$ делит наблюдения на две группы: 1, 2 и 2, 3, вторая функция $D_2(X)$ отделяет наблюдения 2 от всех остальных.

Поле графика разделено дискриминантными функциями на три области: в левой части находятся преимущественно наблюдения первой группы с низкой вероятностью своевременного погашения кредита; в правой части – третьей группы с высокой вероятностью и в нижней части – второй группы со средней вероятностью.



Символы, используемые в территориальной карте

Символ	Грп.	Метка
1	1	Низкая
2	2	Средняя
3	3	Высокая
*		Указывает центр группы

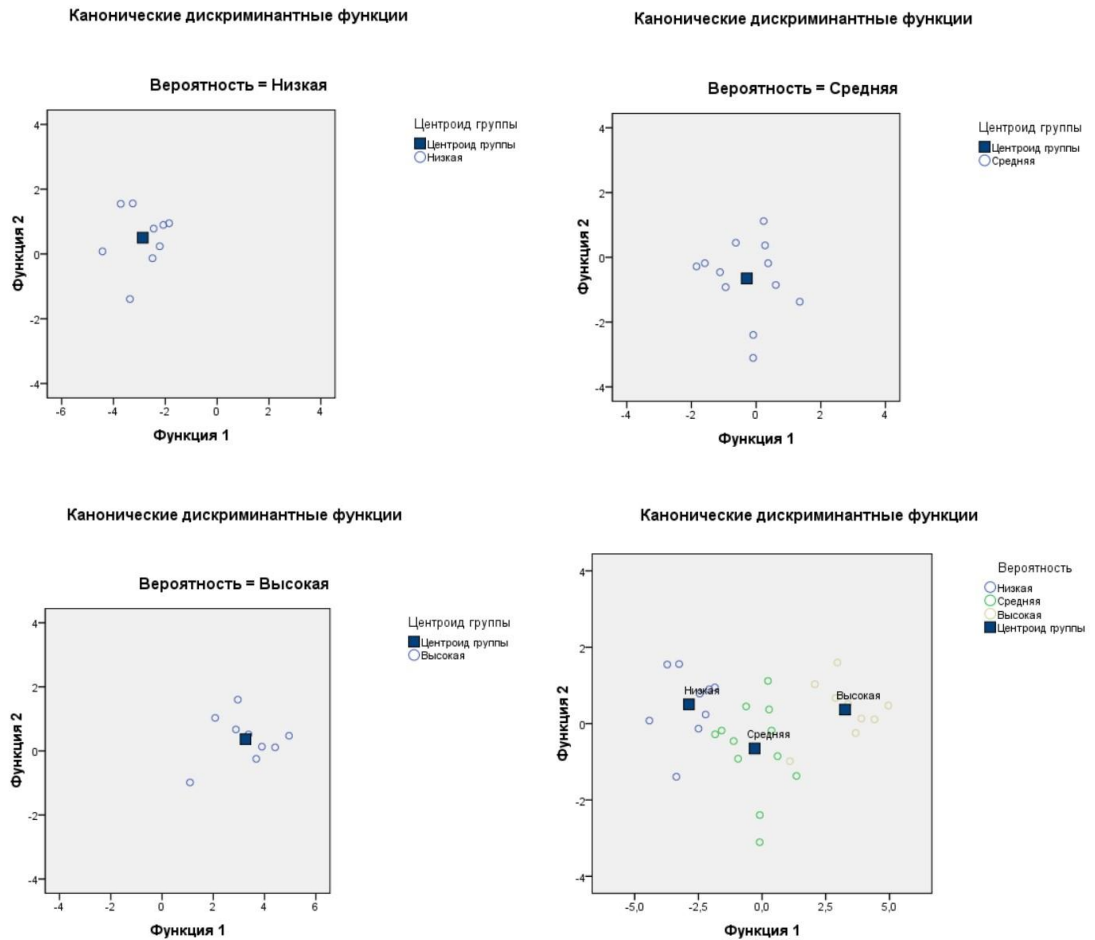
В таблице «Поточечные статистики» размещена информация о фактических (*Actual Group*) и предсказанных (*Predicted Group*) группах для каждого заемщика и соответствующие дискриминантные баллы (*Discriminant Scores*), полученные при подстановке значений переменных в уравнения дискриминантных функций $D_1(X)$ и $D_2(X)$.

Поточечные статистики													
Номер наблюдения	Фактическая группа	Предсказанная группа	Наивероятнейшая группа				Вторая вероятнейшая группа				Дискриминантные баллы		
			P(D=d G=g)		p	ст. св.	P(G=g D=d)	Квадрат расстояния Мхалонобиса до центра	Группа	P(G=g D=d)	Квадрат расстояния Мхалонобиса до центра	Функция 1	Функция 2
			r										
Исходные	1	2	2	2	,721	2	,982	,855	3	,014	8,635	,371	-,183
	2	2	2	2	,779	2	,949	,500	1	,051	5,757	-,942	-,922
	3	2	2	2	,506	2	,973	1,361	3	,017	8,890	,276	,369
	4	3	3	2	,497	2	1,000	1,398	2	,000	22,681	4,412	,111
	5	2	2	2	,202	2	,883	3,200	3	,117	6,666	1,349	-,1372
	6	3	3	2	,235	2	1,000	2,896	2	,000	28,771	4,956	,472
	7	2	2	2	,655	2	,983	,845	3	,016	8,509	,608	-,854
	8	2	2	2	,515	2	,895	1,326	1	,104	5,046	-,627	,449
	9	3	3	2	,794	2	1,000	,462	2	,000	18,130	3,996	,133
	10	1	1	2	,147	2	,943	3,833	2	,057	10,012	-,365	-,1393
	11	3	3	2	,890	2	,996	,232	2	,004	11,800	2,883	,669
	12	3	3	2	,759	2	,999	,552	2	,001	15,874	3,675	-,248
	13	2	2	2	,697	2	,874	,722	1	,126	4,012	-,116	-,459
	14	3	2**	3	,365	2	,927	2,018	3	,073	6,518	1,092	-,966
	15	1	1	2	,275	2	,999	2,585	2	,001	17,638	-,424	,079
	16	1	1	2	,405	2	,999	1,810	2	,001	16,613	-,3720	1,548
	17	1	1	2	,880	2	,950	,257	2	,050	6,730	-,2451	,782
	18	2	1**	2	,433	2	,539	1,673	2	,461	2,558	-,1844	-,282
	19	1	1	2	,675	2	,892	,787	2	,108	5,593	-,2078	,895
	20	3	3	2	,397	2	,952	1,847	2	,048	8,404	2,072	1,031
	21	1	1	2	,778	2	,847	,503	2	,153	4,502	-,2216	,236
	22	2	2	2	,182	2	,944	3,403	3	,030	9,720	,232	1,118
	23	3	3	2	,447	2	,999	1,612	2	,001	15,605	2,957	1,600
	24	1	1	2	,531	2	,997	1,265	2	,003	13,699	-,3257	1,560
	25	3	3	2	,982	2	,999	,036	2	,001	14,788	3,376	,514
	26	2	2	2	,385	2	,598	1,907	1	,402	2,122	-,1587	-,182
	27	1	1	2	,539	2	,832	1,238	2	,168	5,007	-,1854	,948
	28	1	1	2	,761	2	,882	,547	2	,118	5,137	-,2495	-,133
	29	2	2	2	,048	2	,999	6,064	1	,000	20,750	-,094	-,3106
	30	2	2	2	,214	2	,999	3,083	1	,001	16,154	-,089	-,2396
Кросс-проверенные*	1	2	2	6	,554	6	,963	4,919	3	,031	11,239		
	2	2	2	6	,244	6	,866	7,923	1	,134	11,073		
	3	2	2	6	,074	6	,866	11,521	3	,111	15,058		
	4	3	3	6	,853	6	1,000	2,637	2	,000	23,888		
	5	2	2	6	,225	6	,831	8,188	3	,369	8,687		
	6	3	3	6	,178	6	1,000	8,921	2	,000	37,535		
	7	2	2	6	,277	6	,944	7,505	3	,055	12,606		
	8	2	2	6	,437	6	,783	5,878	1	,216	7,881		
	9	3	3	6	,459	6	1,000	5,686	2	,000	22,461		
	10	1	1	6	,047	6	,729	12,743	2	,271	15,297		
	11	3	3	6	,532	6	,990	5,093	2	,010	14,869		
	12	3	3	6	,805	6	,999	3,033	2	,001	17,355		
	13	2	2	6	,870	6	,794	4,046	1	,206	6,188		
	14	3	2**	6	,466	6	,999	5,627	1	,000	21,429		
	15	1	1	6	,212	6	,999	8,381	2	,001	23,699		
	16	1	1	6	,188	6	,999	8,755	2	,001	23,731		
	17	1	1	6	,690	6	,912	3,901	2	,088	9,146		
	18	2	1**	6	,529	6	,852	5,114	2	,148	9,190		
	19	1	1	6	,252	6	,702	7,820	2	,298	10,112		
	20	3	3	6	,243	6	,807	7,939	2	,193	11,378		
	21	1	1	6	,907	6	,796	2,129	2	,204	5,425		
	22	2	3**	6	,003	6	,800	19,944	2	,309	21,845		
	23	3	3	6	,019	6	,993	15,112	2	,007	25,605		
	24	1	1	6	,000	6	,994	25,041	2	,006	35,769		
	25	3	3	6	,965	6	,999	1,419	2	,001	15,394		
	26	2	1**	6	,886	6	,592	3,929	2	,408	5,248		
	27	1	1	6	,675	6	,717	4,013	2	,283	6,446		
	28	1	1	6	,103	6	,574	10,554	2	,426	11,730		
	29	2	2	6	,004	6	1,000	19,379	3	,000	35,465		
	30	2	2	6	,081	6	,998	11,247	1	,001	24,317		
Для исходных данных квадрат расстояния Мхалонобиса вычисляется по канонической функции. Для кросс-проверенных данных квадрат расстояния Мхалонобиса вычисляется по наблюдениям.													
** Неправильно классифицированное наблюдение													
а. Кросс-проверка проводится только для наблюдений в анализе. При кросс-проверке каждое наблюдение классифицируется функциями, выведенными по всем наблюдениям, за исключением его самого.													

На рисунках с 1го по 3ий ниже отражено расположение заемщиков каждой из трех групп на плоскости двух дискриминантных функций $D_1(X)$ и $D_2(X)$. По этим графикам можно проводить детальный анализ вероятностей погашения кредита внутри каждой группы, судить о характере распределения заемщиков и оценивать степень их удаленности от соответствующего центроида.

Кроме того, на 4ом рисунке в той же системе координат приведен объединенный график распределения всех групп заемщиков вместе со своими центроидами; его можно использовать для проведения сравнительного визуального анализа характера взаимного расположения групп заемщиков банка с разными вероятностями погашения кредита. В левой части графика расположены заемщики с низкой вероятностью погашения кредита, в правой – с высокой, а в средней части – со средней вероятностью. Поскольку по результатам расчета вторая дискриминантная

функция $D_2(X)$ оказалась незначима, то различия координат центроидов по этой оси незначительны. Этот факт подтверждается картой восприятия, которая была расположена выше.



Данные таблицы «Результаты классификации» свидетельствуют о том, что для 93,3% исходных и перекрестно-проверяемых 86,7% сгруппированных наблюдений классификация проведена корректно, высокая точность достигнута в каждой из групп, но в первой она максимальная – 100%, а в третьей несколько ниже – 88,9%.

Результаты классификации^{b,c}

			Предсказанная принадлежность к группе			Итого
			Низкая	Средняя	Высокая	
Исходные	Частота	Низкая	9	0	0	9
		Средняя	1	11	0	12
		Высокая	0	1	8	9
	%	Низкая	100,0	,0	,0	100,0
		Средняя	8,3	91,7	,0	100,0
		Высокая	,0	11,1	88,9	100,0
Кросс-проверенные ^a	Частота	Низкая	9	0	0	9
		Средняя	2	9	1	12
		Высокая	0	1	8	9
	%	Низкая	100,0	,0	,0	100,0
		Средняя	16,7	75,0	8,3	100,0
		Высокая	,0	11,1	88,9	100,0

a. Кросс-проверка проводится только для наблюдений в анализе. При кросс-проверке каждое наблюдение классифицируется функциями, выведенными по всем наблюдениям, за исключением его самого.

b. 93,3% исходных сгруппированных наблюдений классифицировано правильно.

c. 86,7% перекрестно-проверяемых сгруппированных наблюдений классифицировано правильно.

Вывод: приобрёл практические навыки применения дискриминантного анализа для решения конкретных задач с использованием статистического пакета SPSS.