jPrediction - бинарный классификатор для машинного обучения

Автор Reshetov вкл Tuesday, 28 June 2016

Оглавление

Основные характеристики

1. Предназначение: бинарная классификации объектов по их признакам (предикторам)
2. Устройство: нейросеть прямого распространения с структуированным скрытым слоем
3. Метод обучения: критерий минимакса по теореме о минимаксе фон Неймана-Моргенштерна. Для реализации используется библиотека libVMR
4. Достоинства: высокая обобщающая способность и автоматическая редукция неинформативных предикторов
5. Недостатки: значительное потребление вычислительных ресурсов по расходу памяти и продолжительности вычислений. Каждый дополнительный предиктор увеличивает потребление вдвое.
6. Лицензия: GNU GPL Version 3

Запуск jPrediction

Чтобы запустить jPrediction как приложение, необходимо загрузить его на компьютер по этой ссылке: jPrediction.jar

После того, как файл загружен на компьютер, его можно запустить двойным кликом. Если приложение не запускается, то скорее всего на Вашем компьютере либо вообще не установлена, либо установлена старая версия Java. Обновить версию можно на сайте: www.java.com

Исходные файлы приложения можно скачать в виде архива, по ссылке:jPrediction_open_source.zip. Этот архив можно импортировать в различные среды разработки Java, например, в Eclipse

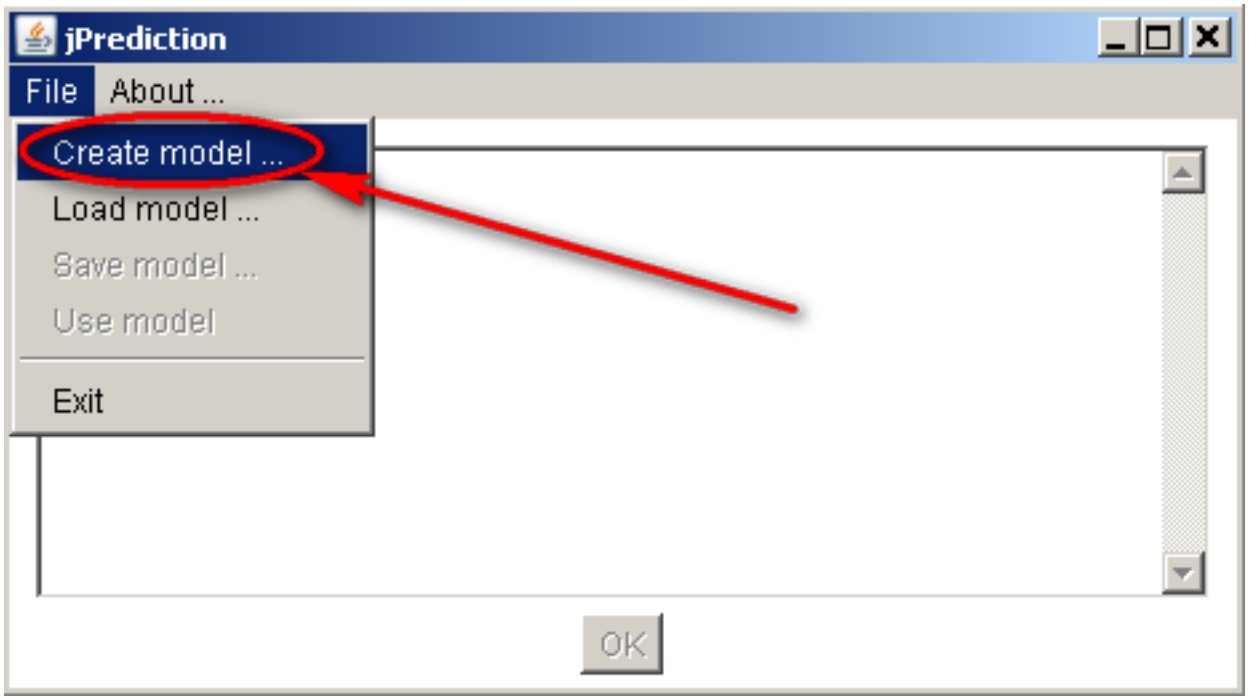
Как создать математическую модель бинарного классификатора в jPrediction

Для создания математических моделей необходимо собрать данные и записать их в файл формата CSV - генеральную выборку.

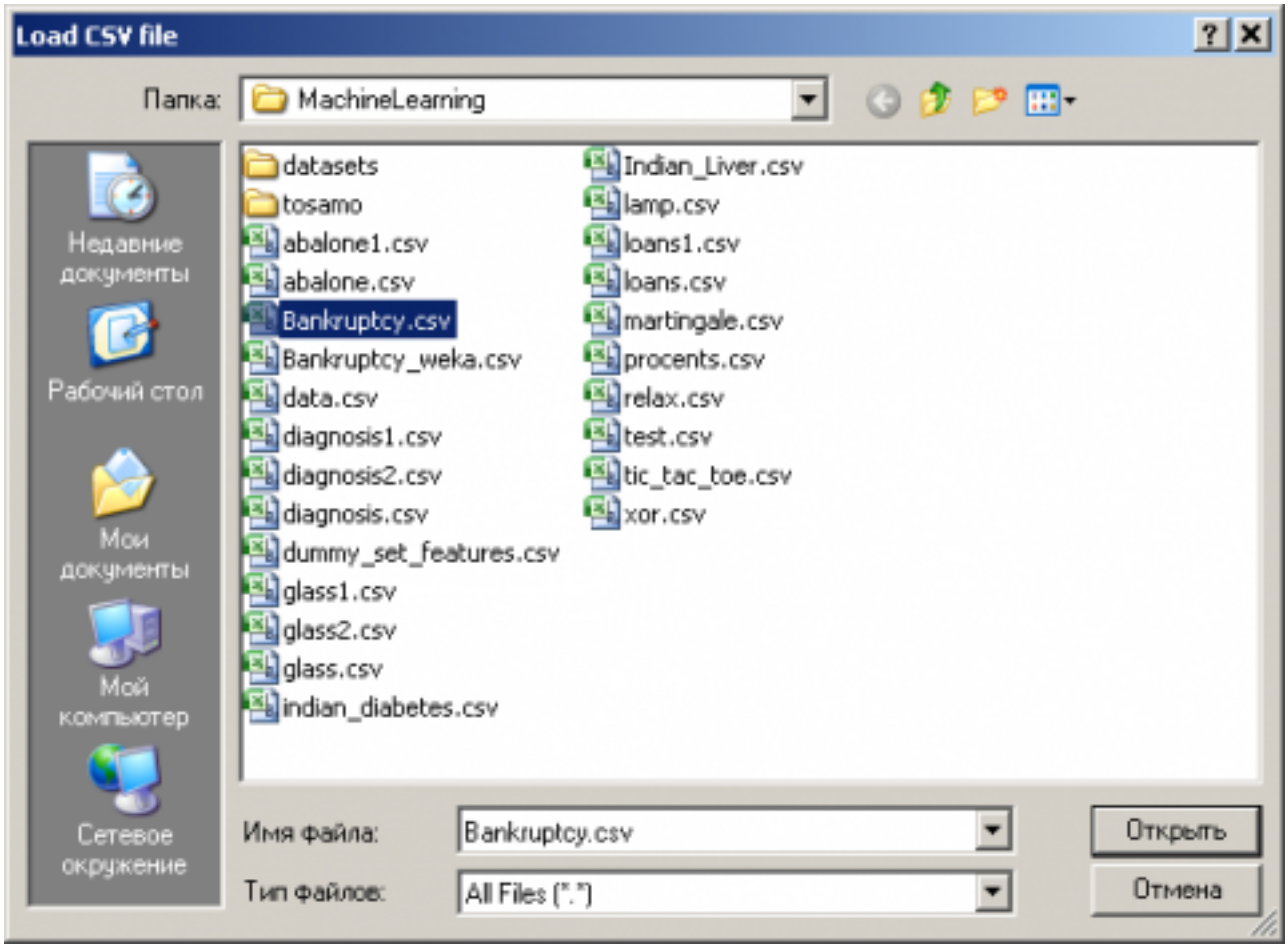
В качестве примера мы возьмём готовую выборку Bankruptcy.csv (прогнозирование банкротств юридических лиц), которая в свою очередь была взята из репозитория

https://archive.ics.uci.edu/ml/datasets/Qualitative_Bankruptcy

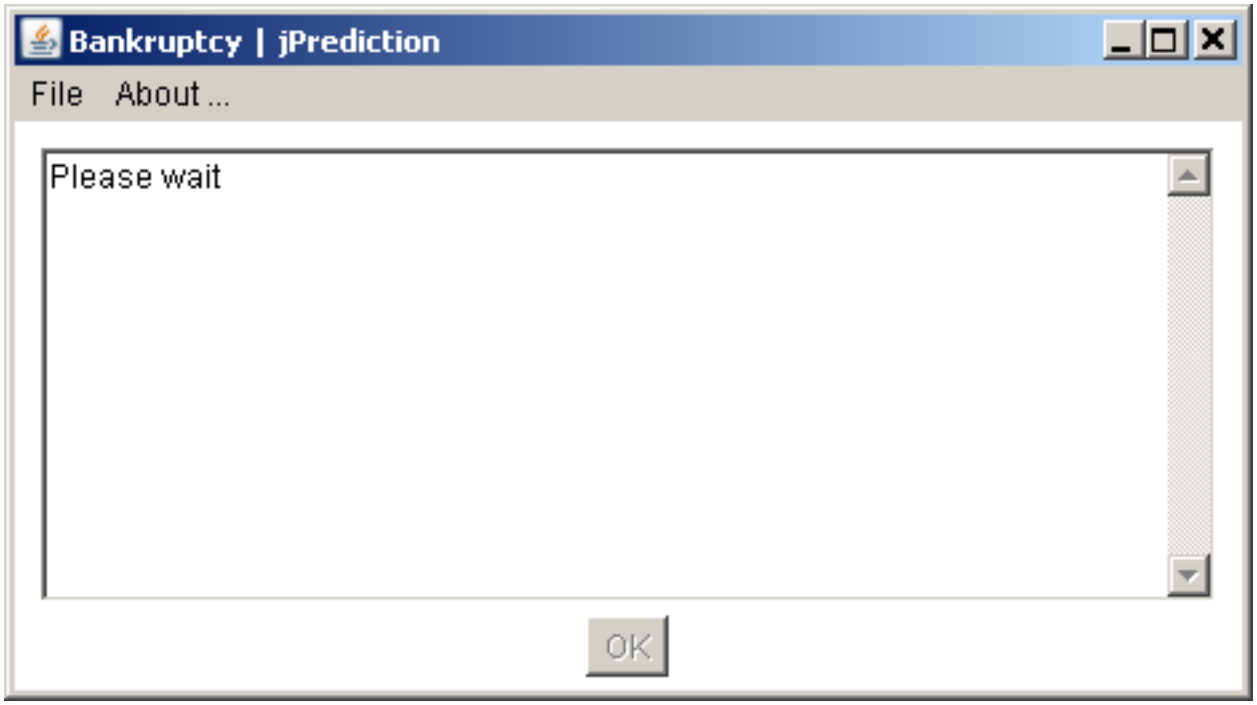
После чего нужно запустить jPrediction и выбрать пункт меню File>Create model ...:



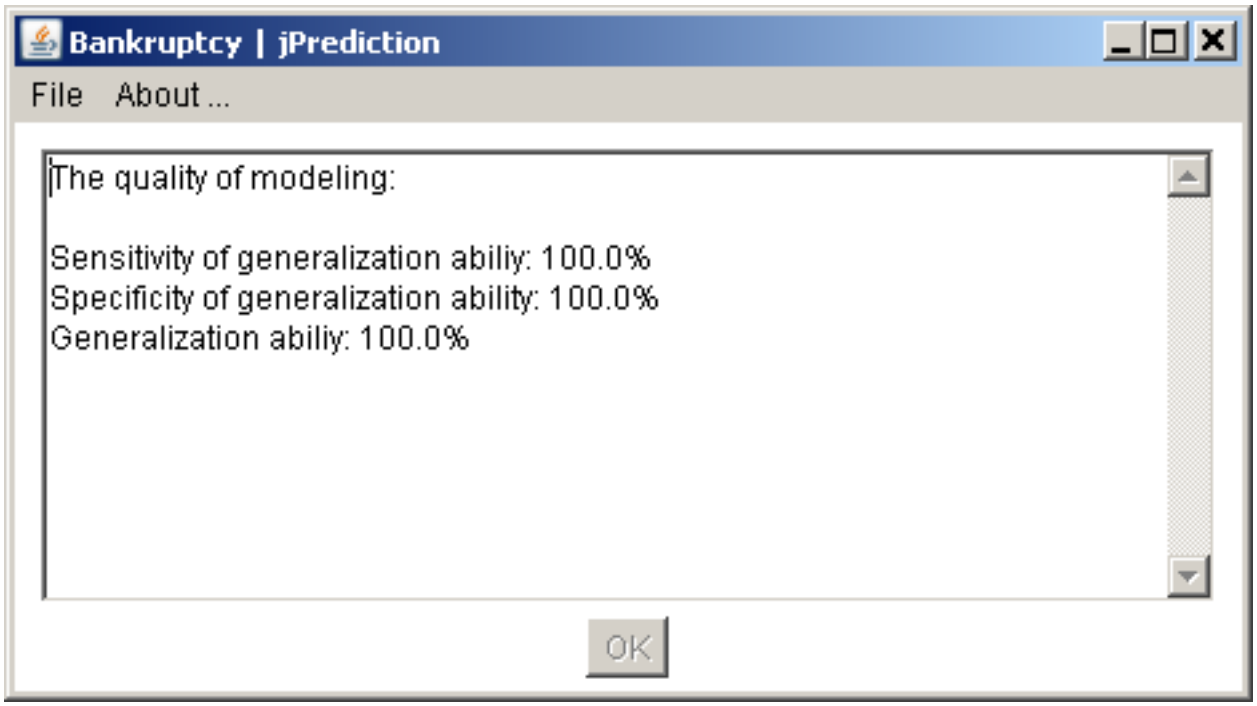
В появившемся диалоговом меню выбираем файл формата CSV с данными:



После чего выборка случайным образом разделяется на две части: обучающую и тестовую. Далее автоматически запускается процесс обучения на обучающей выборке и надпись: "Please wait" (Подождите, пожалуйста):



После завершения обучения, полученная математическая модель проверяется на тестовой части выборки (обобщающая спосособность) и результаты такой проверки выводятся в виде отчёта:

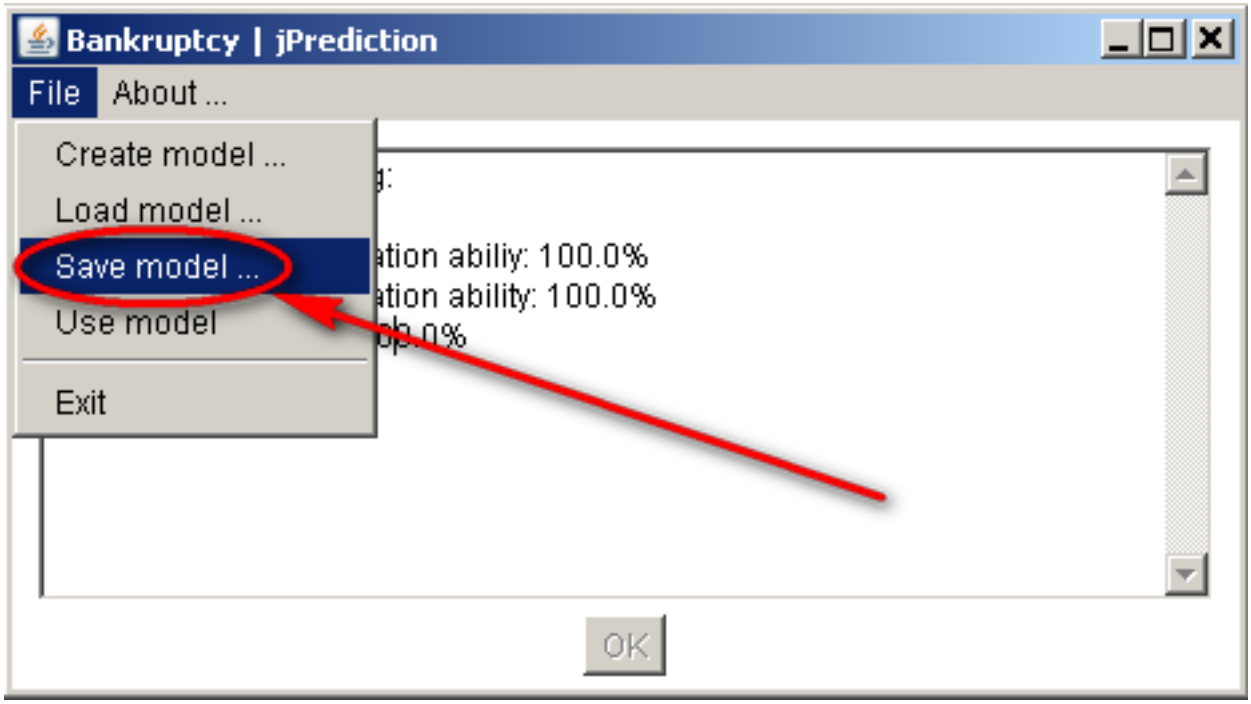


Здесь:

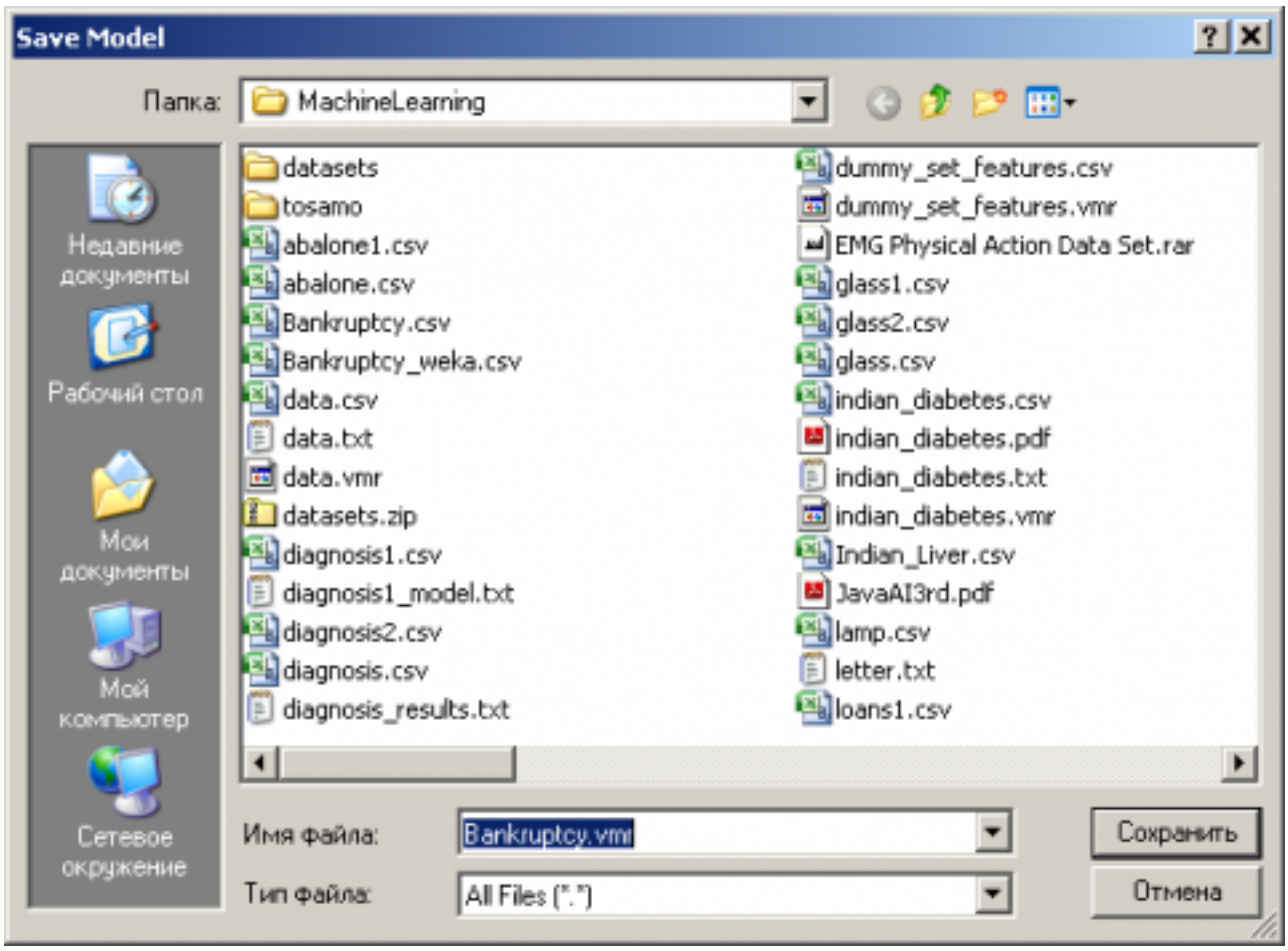
- Sensitivity - это чувствительность модели в процентах
- Specificity - специфичность модели в процентах

Сохранение модели в файл

Чтобы каждый раз не создавать математическую модель заново, обучая бинарный классификатор, её можно записать в файл формата VMR. Для этого нужно выбрать пункт меню File>Save model ...:



В диалоговом окне уже указано имя файла, которое совпадает с именем загруженного CSV, но имеет расширение VMR. Поэтому можно сразу же нажать кнопку "Сохранить", если нет никакой необходимости менять имя файла:

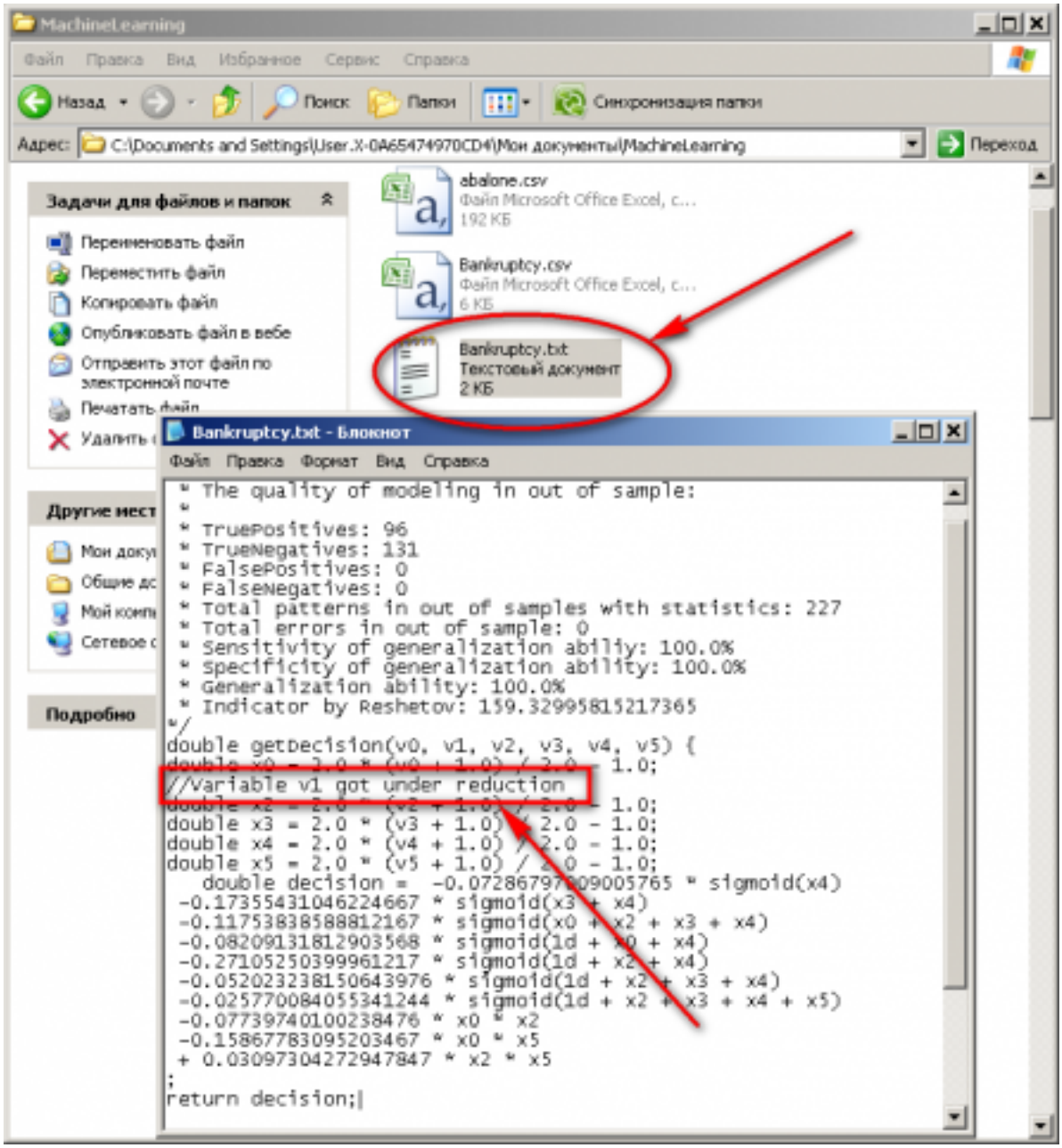


Редукция - удаление неинформативных признаков из модели

После того как модель создана и сохранена на диск, создаются два файла:

1. Файл модели для использования, который можно впоследствии загрузить с расширением VMR
2. Файл кода модели на Java с детальным отчётом с расширением TXT

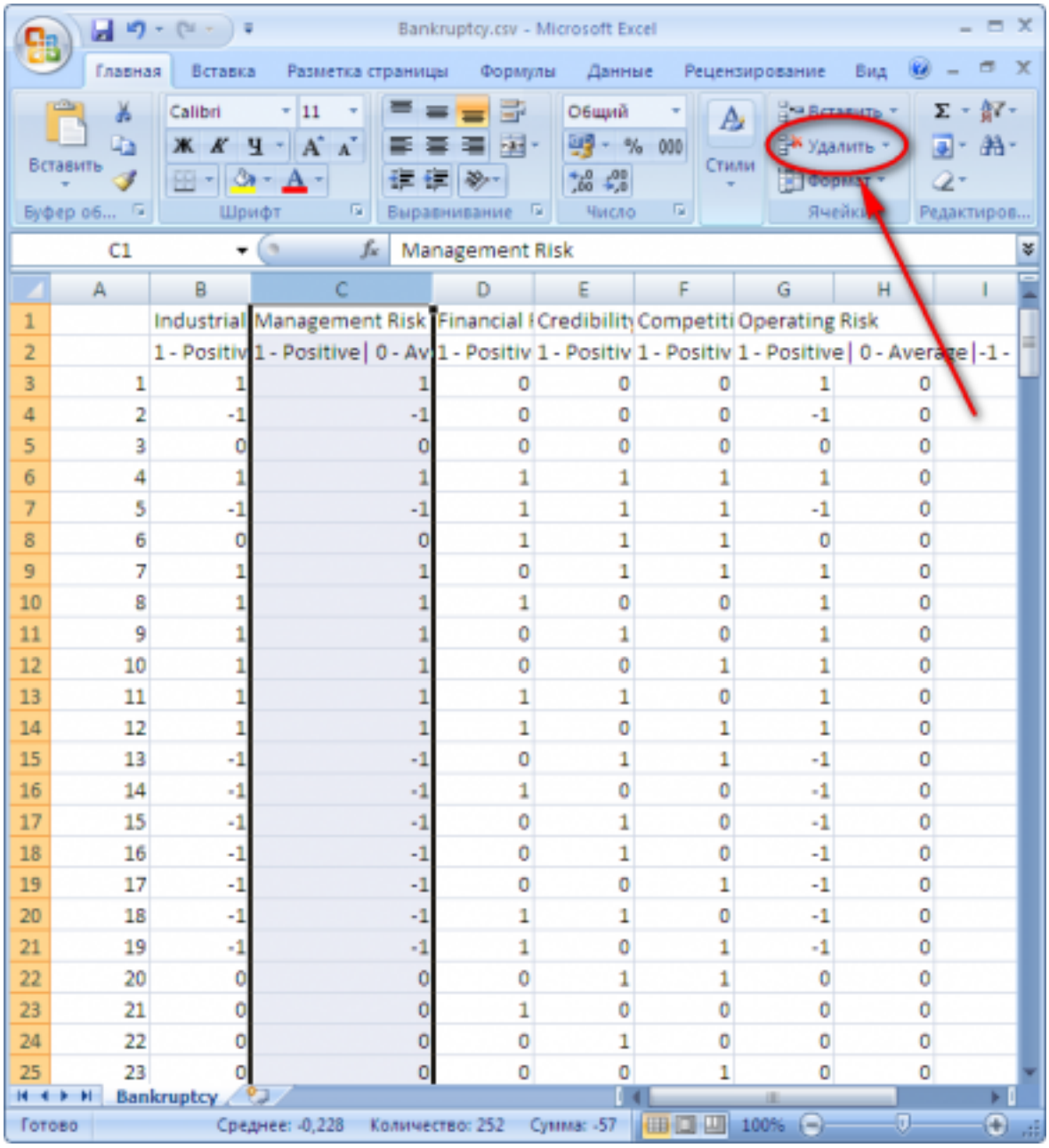
Нас будет интересовать файл с отчётом, имеющий расширение TXT. Откроем его для просмотра.



В этом файле имеется запись: "Variable v1 got under reduction", т.е. переменная с идентификатором v1 редуцирована из модели. В чём нетрудно убедиться, если посмотреть код модели на Java. Там отсутствует переменная x1. Соответственно значение этой переменной нигде не используется. Однако, в самой модели её идентификатор остался и его значение нужно будет всякий раз вводить при применении модели для классификации, что нежелательно.

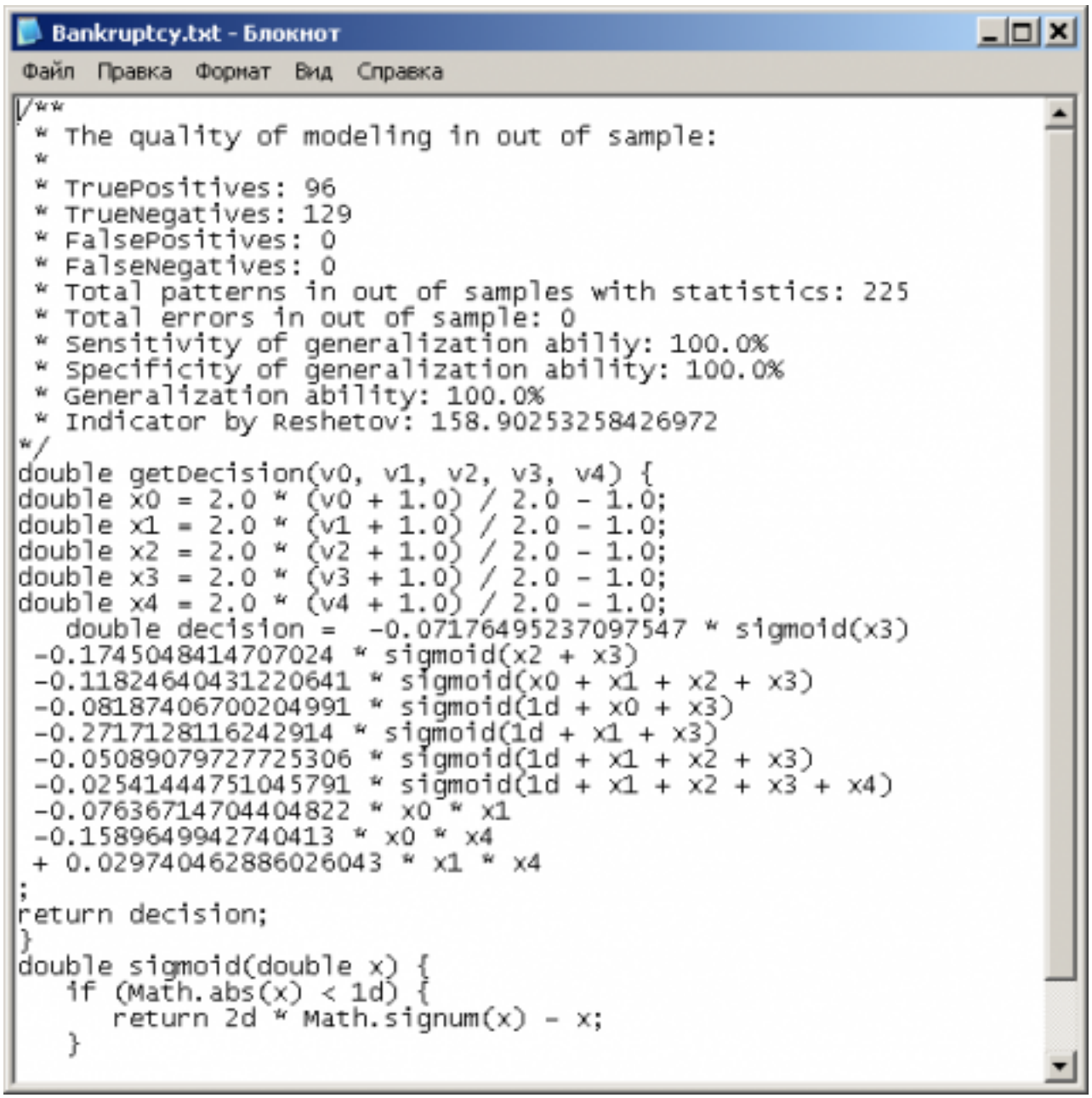
Поэтому нам нужно удалить лишний предиктор из выборки (файла CSV) и переобучить модель заново. Прибавляем к номеру переменной двойку и получаем число 3. Т.е. из выборки нужно удалить третью колонку.

Третья колонка у нас называется "Management risk" (Управленческий риск). Удаляем её из электронной таблицы:



	A	B	C	D	E	F	G	H	I
1		Industrial	Management Risk	Financial I	Credibility	Competiti	Operating Risk		
2		1 - Positiv	1 - Positive	0 - Av	1 - Positiv	1 - Positiv	1 - Positiv	1 - Positive	0 - Average	-1 -
3	1	1	1	0	0	0	1	0	
4	2	-1	-1	0	0	0	-1	0	
5	3	0	0	0	0	0	0	0	
6	4	1	1	1	1	1	1	0	
7	5	-1	-1	1	1	1	-1	0	
8	6	0	0	1	1	1	0	0	
9	7	1	1	0	1	1	1	0	
10	8	1	1	1	0	0	1	0	
11	9	1	1	0	1	0	1	0	
12	10	1	1	0	0	1	1	0	
13	11	1	1	1	1	0	1	0	
14	12	1	1	1	0	1	1	0	
15	13	-1	-1	0	1	1	-1	0	
16	14	-1	-1	1	0	0	-1	0	
17	15	-1	-1	0	1	0	-1	0	
18	16	-1	-1	0	1	0	-1	0	
19	17	-1	-1	0	0	1	-1	0	
20	18	-1	-1	1	1	0	-1	0	
21	19	-1	-1	1	0	1	-1	0	
22	20	0	0	0	1	1	0	0	
23	21	0	0	1	0	0	0	0	
24	22	0	0	0	1	0	0	0	
25	23	0	0	0	0	1	0	0	

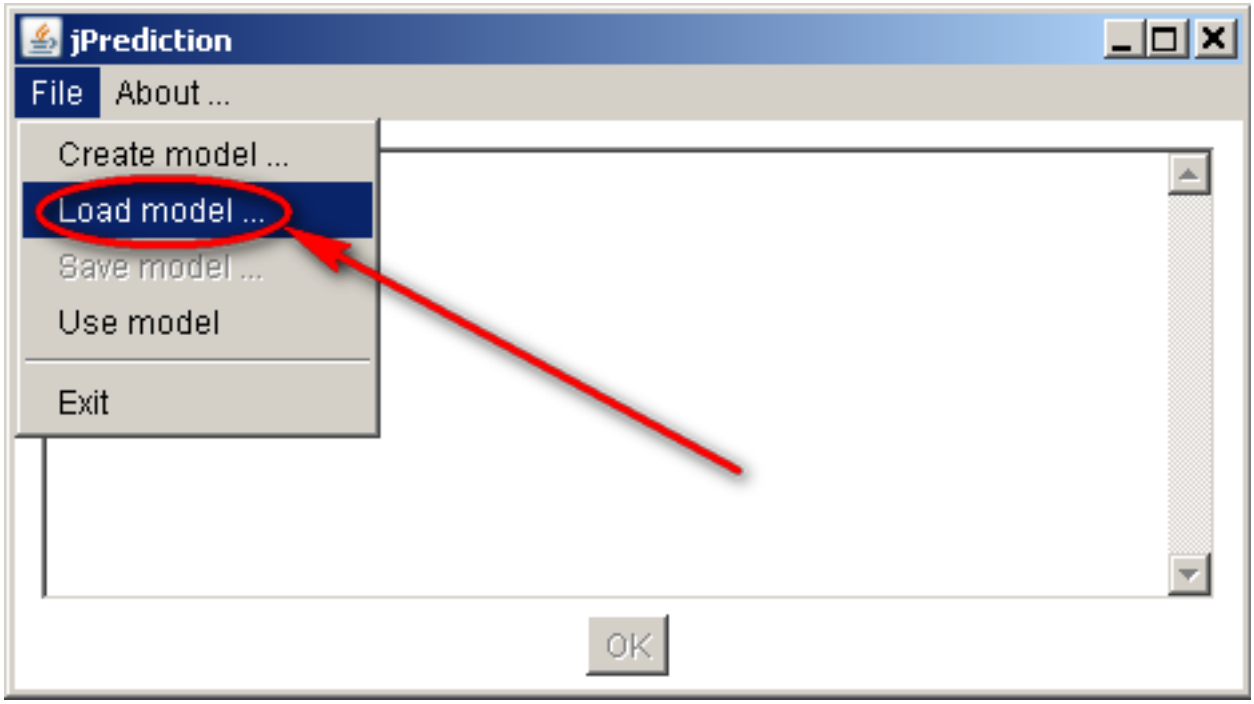
Переобучаем модель на изменённой выборке и получаем такие результаты:



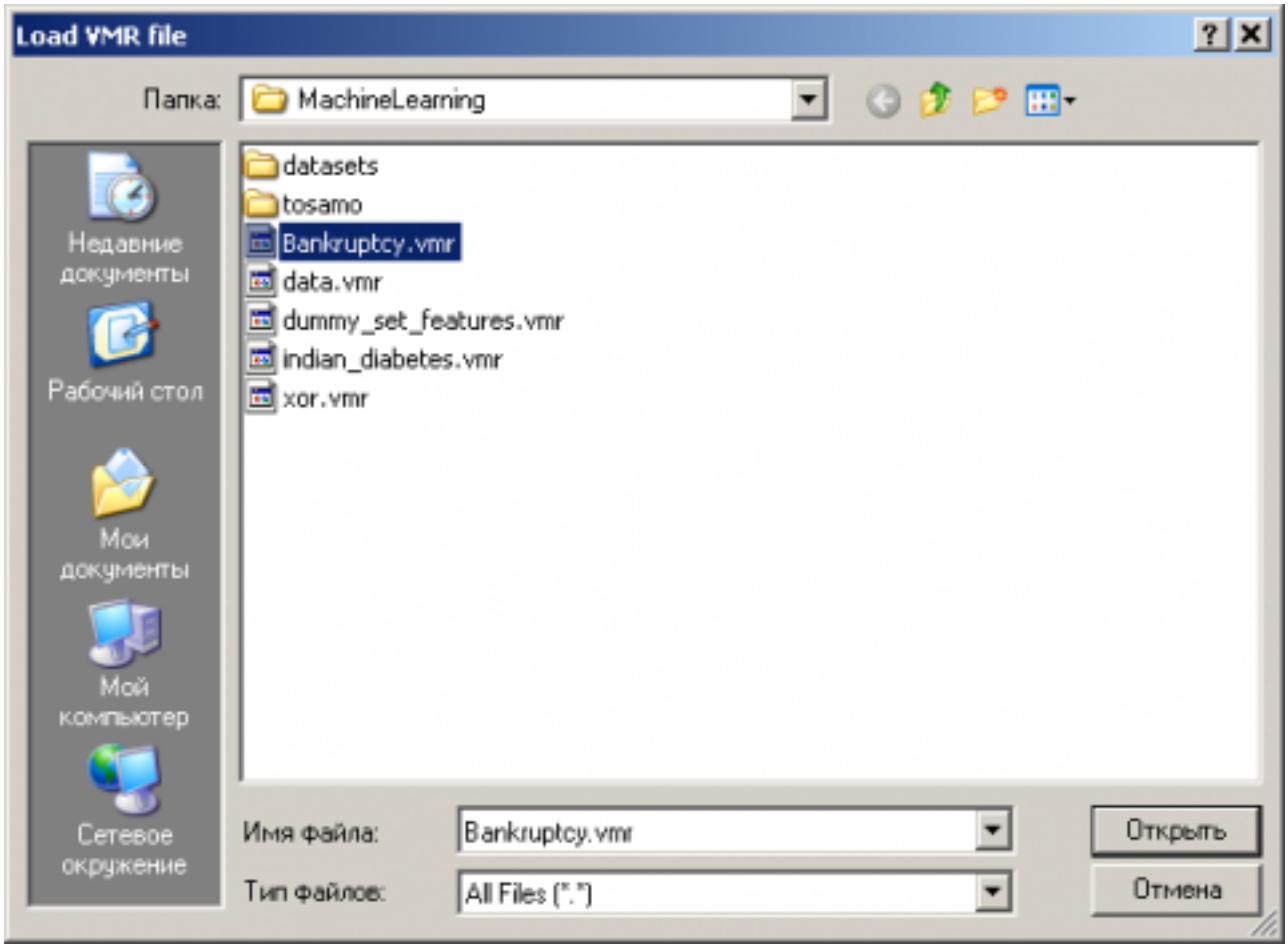
Теперь количество предикторов в нашей модели стало на один меньше.

Загрузка и использование модели для классификации объектов

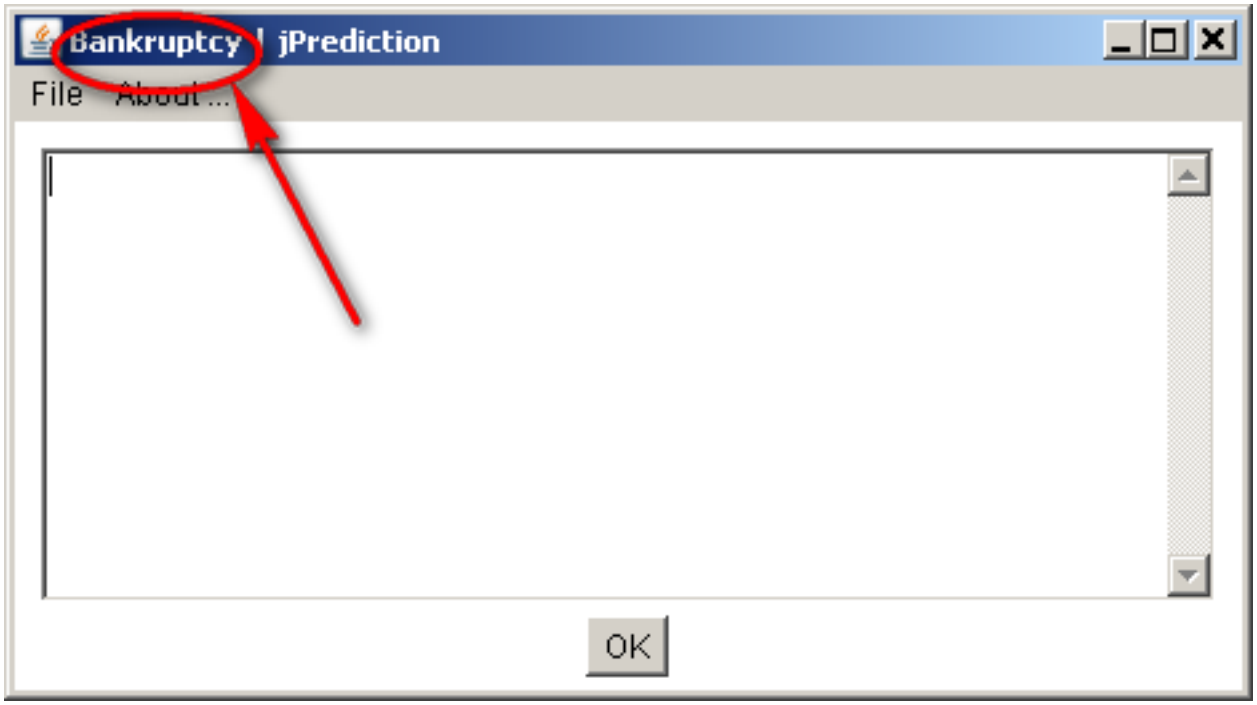
Если впоследствии понадобится загрузить модель для использования, то для этого можно выбрать пункт меню File>Load model ...:



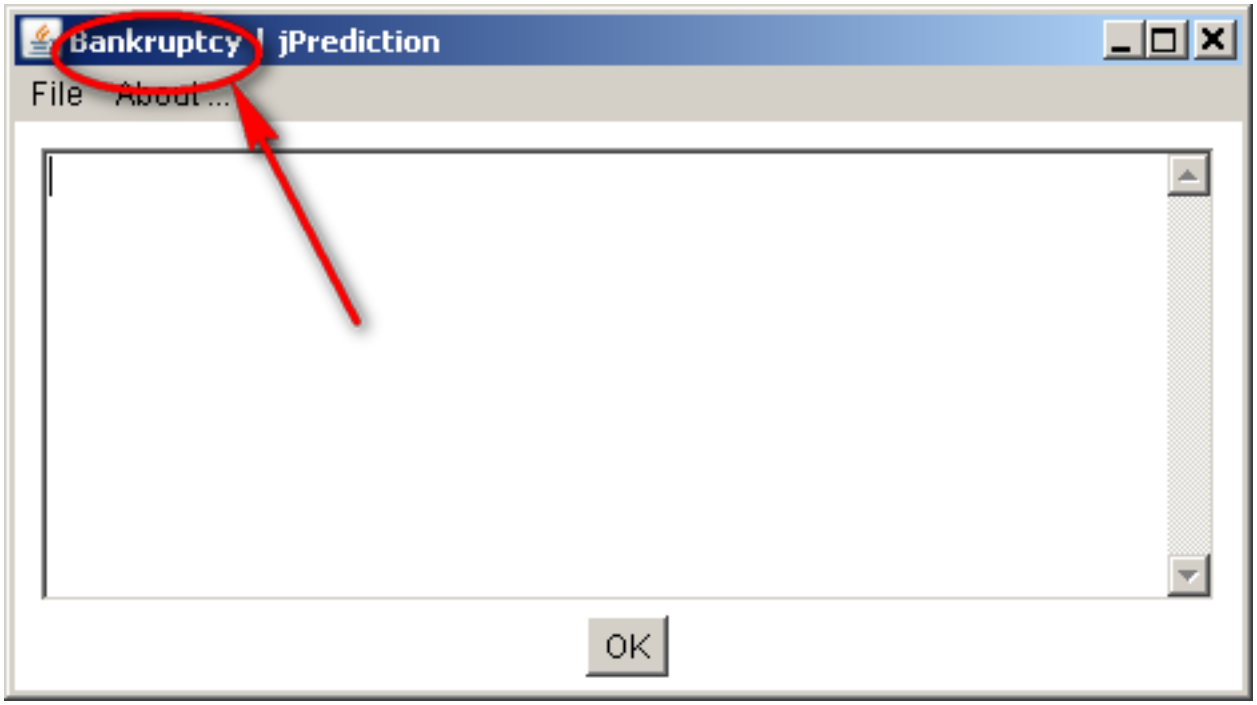
В появившемся диалоговом окне выбираем нужный нам файл модели и нажимаем кнопку "Открыть":



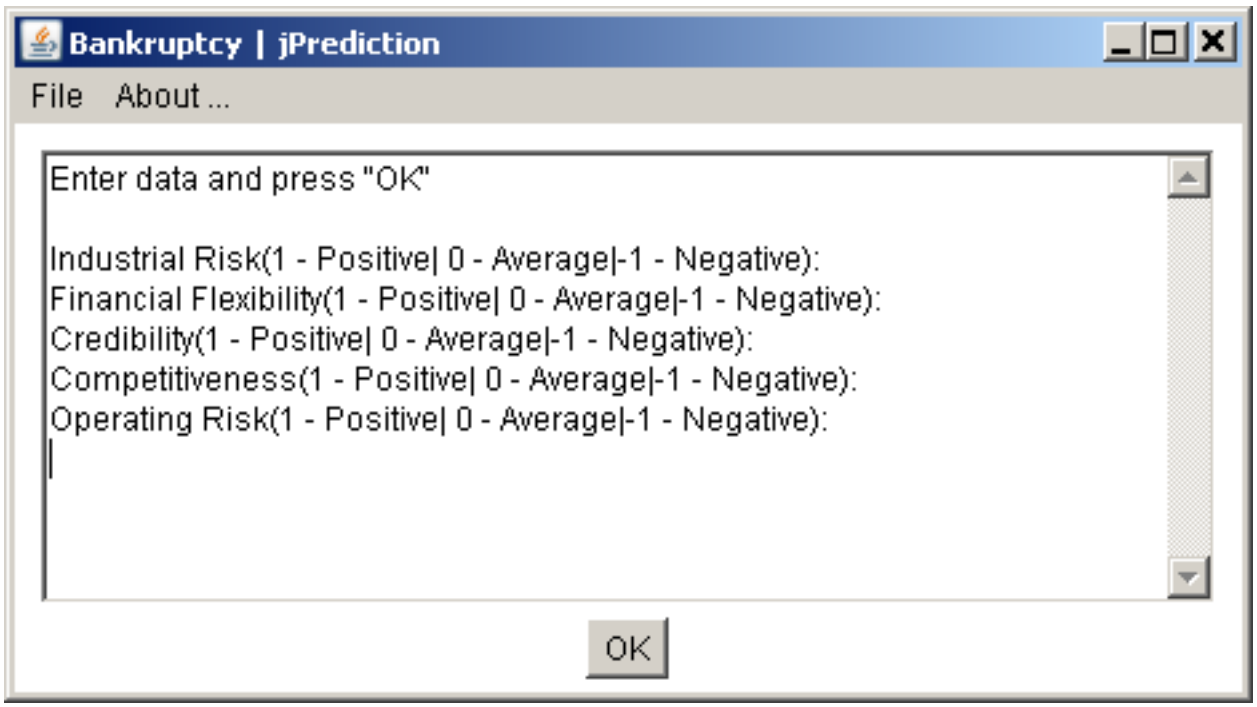
В заголовке окна появилось название загруженной модели:



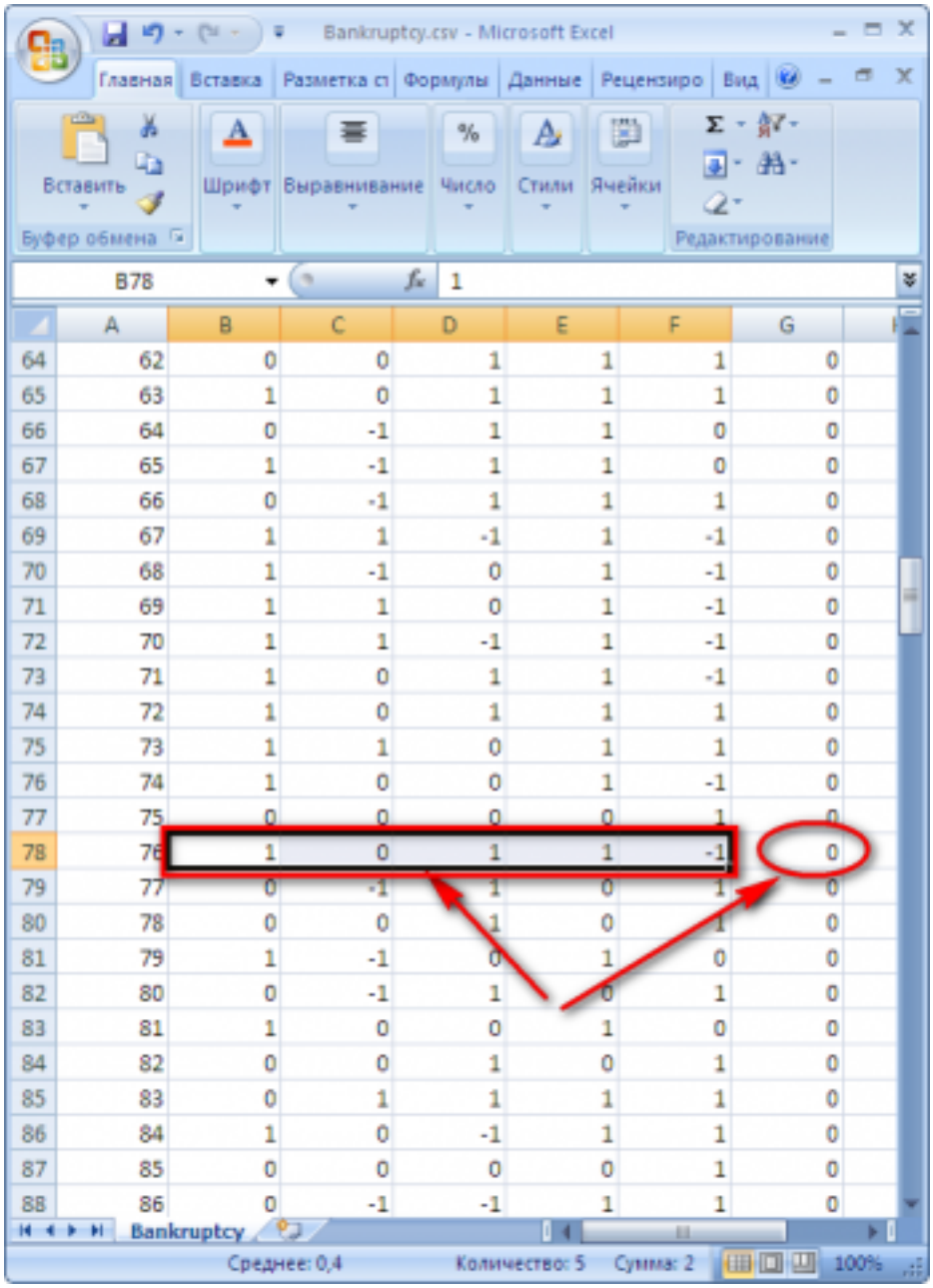
Чтобы классифицировать объекты, нужно выбрать пункт меню File>Use Model:



Текстовое окно заполнится названиями предикторов и примечаниями для них в скобках:

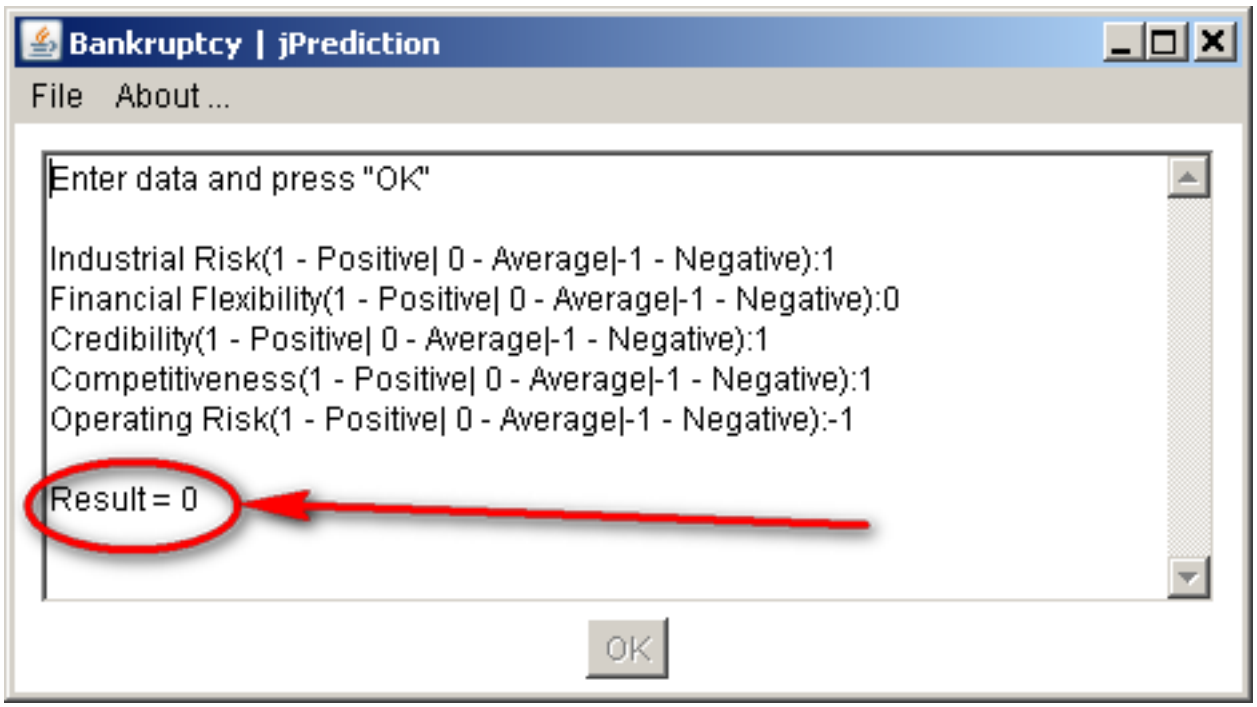


Попробуем ввести данные для какого нибудь примера из выборки:



	A	B	C	D	E	F	G
64	62	0	0	1	1	1	0
65	63	1	0	1	1	1	0
66	64	0	-1	1	1	0	0
67	65	1	-1	1	1	0	0
68	66	0	-1	1	1	1	0
69	67	1	1	-1	1	-1	0
70	68	1	-1	0	1	-1	0
71	69	1	1	0	1	-1	0
72	70	1	1	-1	1	-1	0
73	71	1	0	1	1	-1	0
74	72	1	0	1	1	1	0
75	73	1	1	0	1	1	0
76	74	1	0	0	1	-1	0
77	75	0	0	0	0	1	0
78	76	1	0	1	1	-1	0
79	77	0	-1	1	0	1	0
80	78	0	0	1	0	1	0
81	79	1	-1	0	1	0	0
82	80	0	-1	1	0	1	0
83	81	1	0	0	1	0	0
84	82	0	0	1	0	1	0
85	83	0	1	1	1	1	0
86	84	1	0	-1	1	1	0
87	85	0	0	0	0	1	0
88	86	0	-1	-1	1	1	0

После чего нажмём кнопку OK:



И как видим, результат у нас правильный, что неудивительно при 100% обобщающей способности.

Приложение:

Дополнительные выборки для бинарной классификации

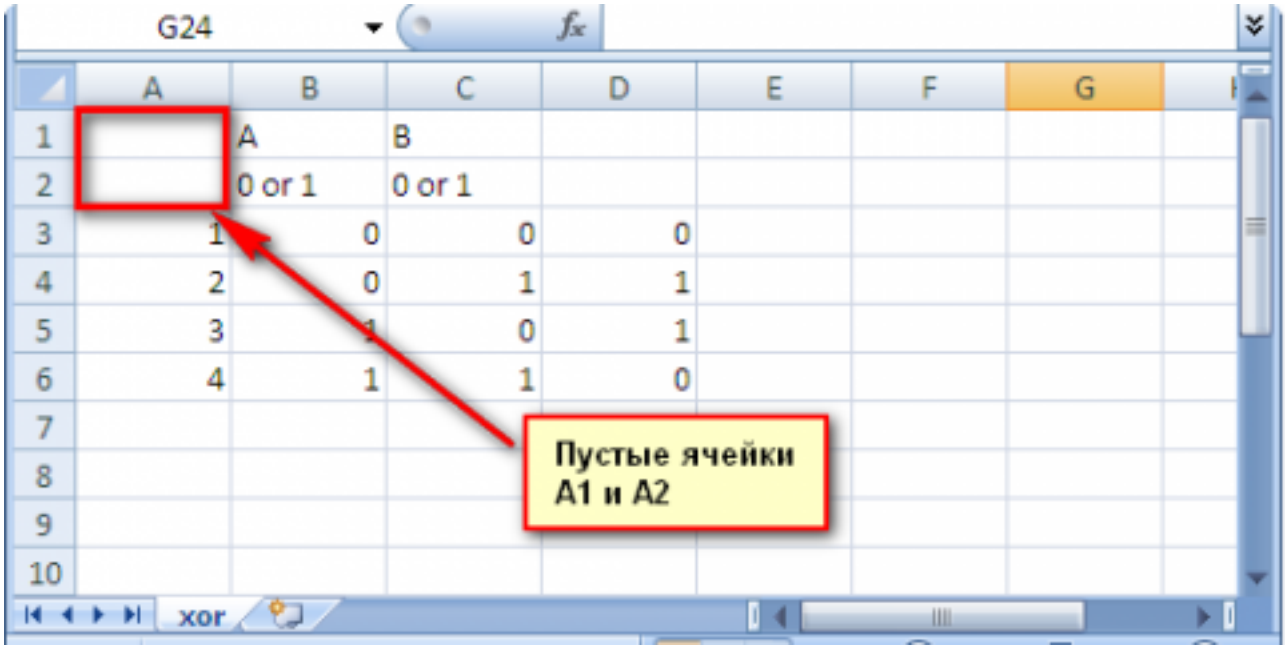
Возможно у кого-то возникнет желание применить jPrediction для других задач бинарной классификации. Например, для задач из статьи: Применение логистической регрессии в медицине и скоринге. Их можно взять из нижеприведённого списка:

1. Банковский скоринг - loans.csv
2. Диагностика диабета - indian_diabetes.csv

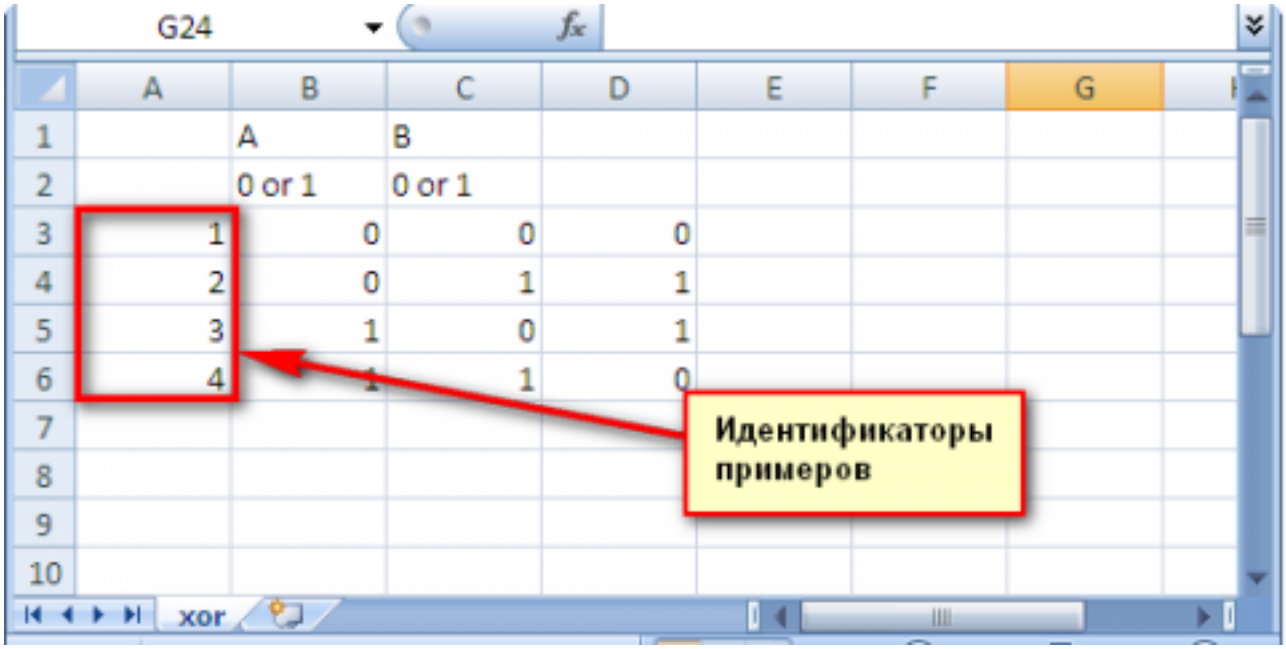
Формат CSV файлов для jPrediction

Рассмотрим задачу двоичного исключающего ИЛИ (XOR) и заполнение электронной таблицы для неё.

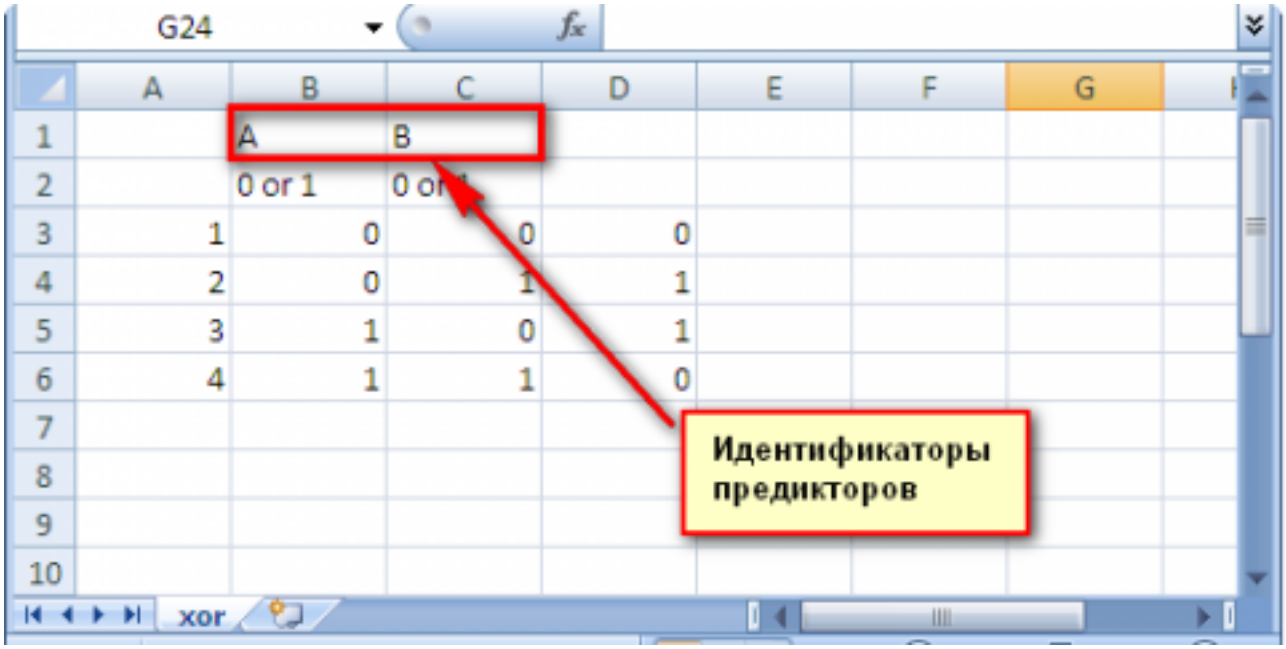
Ячейки A1 и A2 должны быть всегда пустыми:



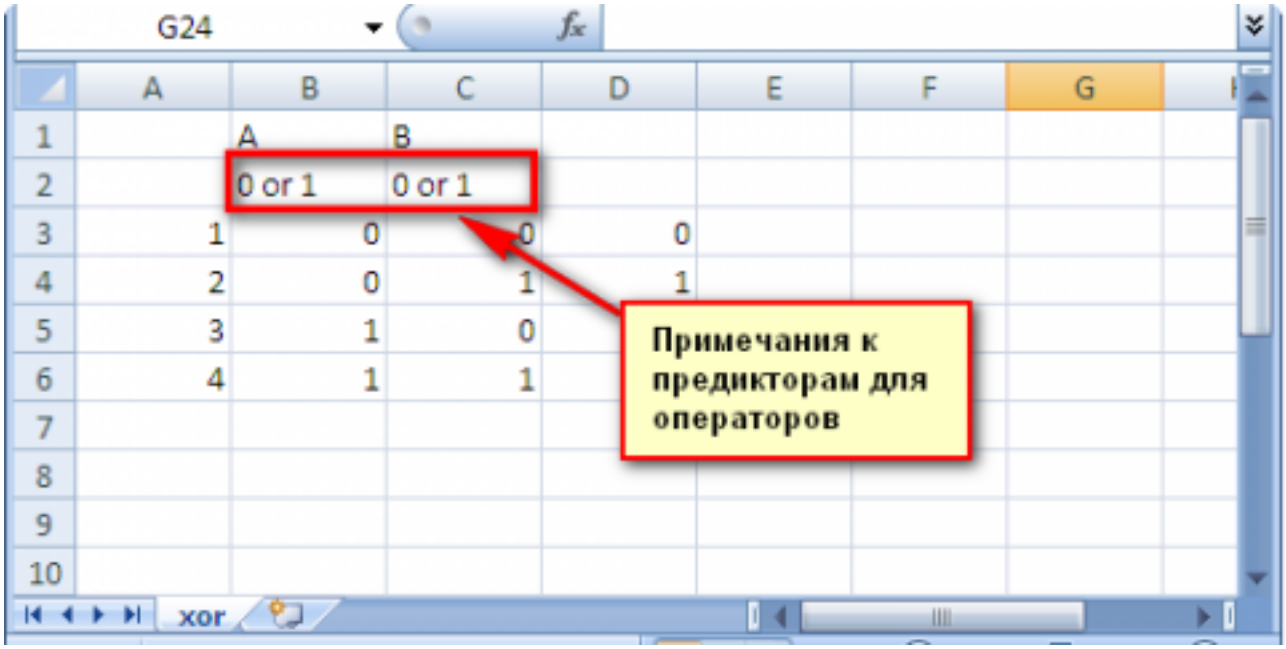
В первой колонке располагаются идентификаторы примеров, которые не применяются для создания математических моделей, но служат для контроля правильности ввода данных согласно журнала исследований:



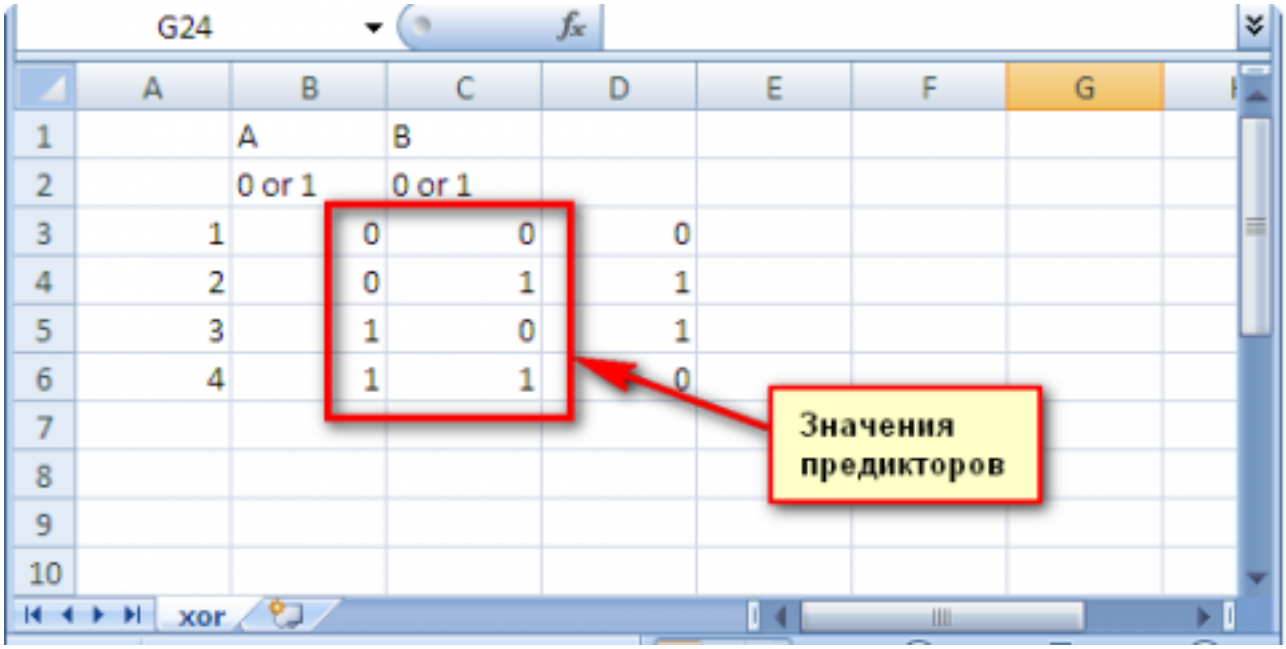
В первой строке выписываются идентификаторы предикторов:



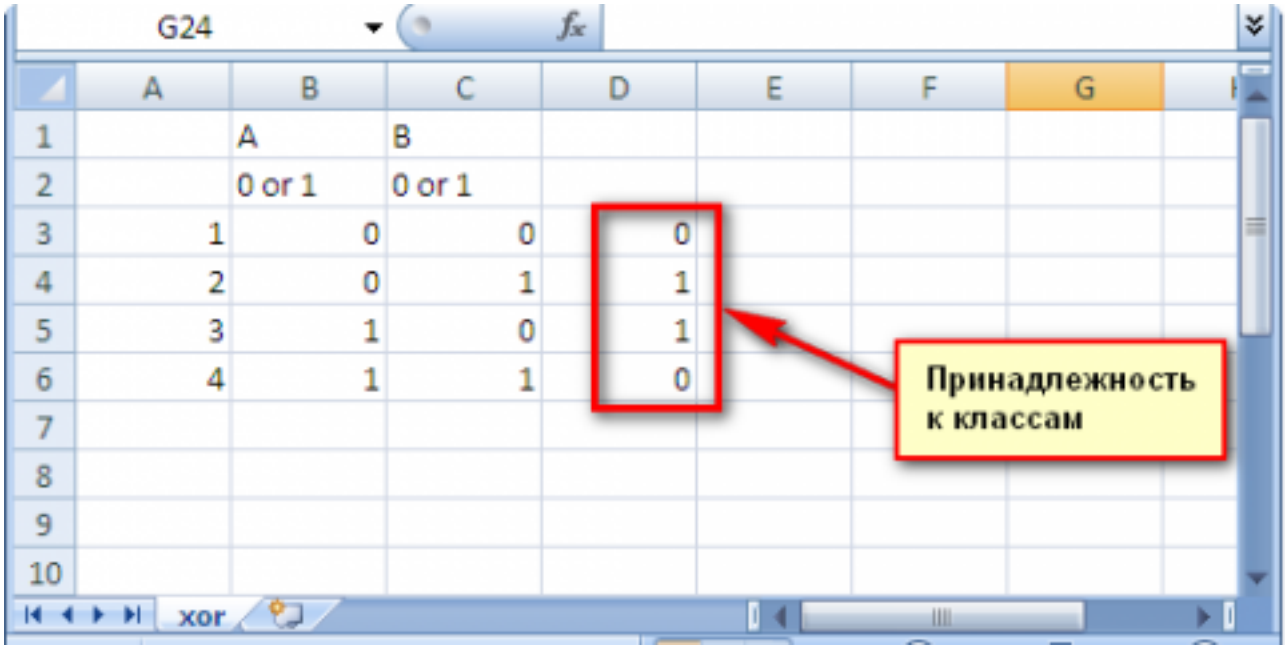
Вторая строка - примечания к предикторам, которые будут отображены при использовании классификатора. Это подсказки для тех, кто будет классифицировать объекты по загруженным моделям:



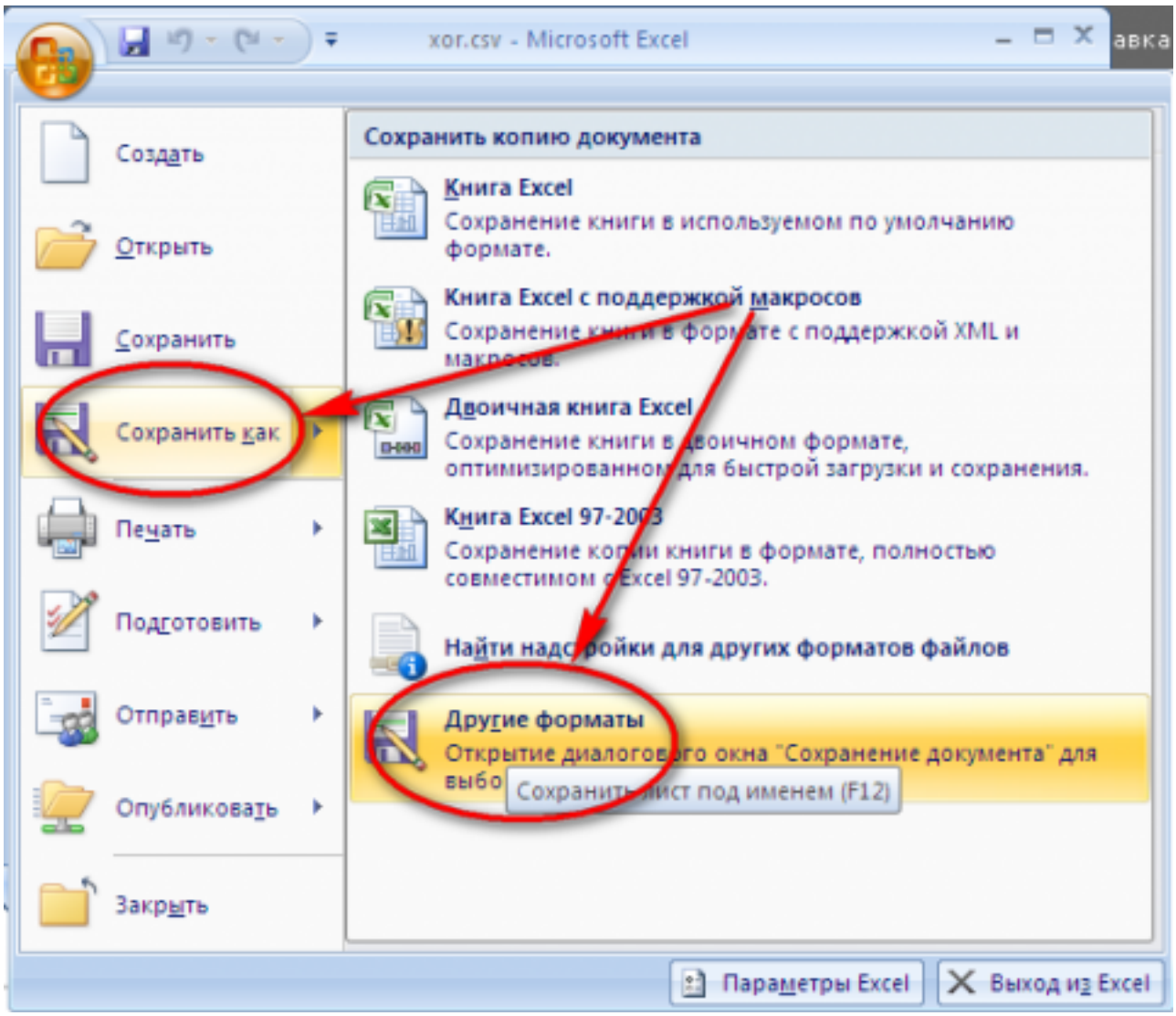
Начиная с третьей строки, кроме первой и последней колонок, располагаются значения предикторов. Все значения предикторов в jPrediction должны быть только числовыми. Категориальные значения недопустимы:

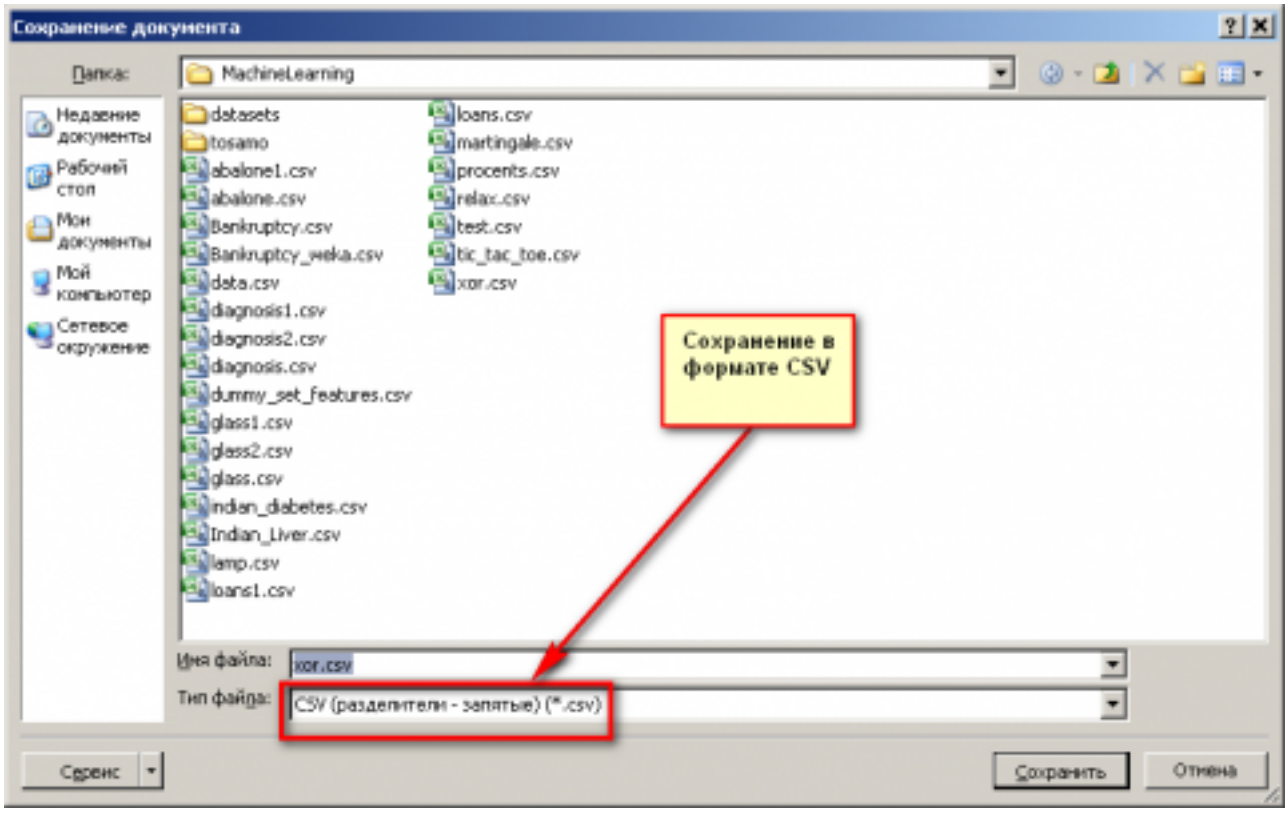


И в последнем столбце указываются значения зависимой переменной в бинарном виде: 0 - принадлежность объектов к одному классу, 1 - принадлежность к другому классу (две верхние ячейки в строках 1 и 2 должны быть пустыми):



Если файл создаётся впервые, то для его сохранения в формате CSV нужно выбрать "Сохранить как ...":


xor.csv - Microsoft Excel
Создать
Открыть
Сохранить
Сохранить как
Печать
Подготовить
Отправить
Опубликовать
Закрыть
Сохранить копию документа
Книга Excel
Сохранение книги в используемом по умолчанию формате.
Книга Excel с поддержкой макросов
Двоичная книга Excel
Сохранение книги в двоичном формате, оптимизированном для быстрой загрузки и сохранения.
Книга Excel 97-2003
Найти надстройки для других форматов файлов
Другие форматы
Сохранить лист под именем (F12)
Параметры Excel
Выход из Excel


Сохранение документа
Папка:
MachineLearning
datasets
tosamo
abalone1.csv
abalone.csv
Bankruptcy.csv
Bankruptcy_учка.csv
data.csv
diagnosis1.csv
diagnosis2.csv
diagnosis.csv
dummy_set_features.csv
glass1.csv
glass2.csv
glass.csv
indian_diabetes.csv
Indian_Liver.csv
lamp.csv
loans1.csv
loans.csv
martingale.csv
procents.csv
relax.csv
test.csv
tic_tac_toe.csv
xor.csv
Сохранение в формате CSV
Имя файла:
xor.csv
Тип файла:
CSV (разделители - запятые) (*.csv)
Сервис
Сохранить
Отмена

Юрий Решетов

Tags: Математика
Язык Русский

Source URL: http://yury-reshetov.com/node/150