***ФГБОУ ВО «Кубанский государственный аграрный университет имени И.Т. Трубилина»,***

***Российская Федерация***

***Мальцева Эмма Эдуардовна, ПИ2102
emmamaltseva0@gmail.com***

**1 Практическое занятие Интеллектуальные технологии и представление знаний. 2021-12-02**

**Заголовок:**

Разработка и анализ модели кредитного скоринга в системе Эйдос

**Резюме:**

В видео демонстрируется процесс создания и анализа модели кредитного скоринга с использованием интеллектуальной системы Эйдос для оценки риска невозврата кредита физическими лицами.

1. **Подготовка и Настройка:**
	* Определены типы шкал: две классификационные (2-3) и описательные (4-10).
	* Запущена система Эйдос, удалены предыдущие приложения.
2. **Ввод и Подготовка Данных:**
	* Данные импортированы из Excel-файла с использованием программного интерфейса 2-3-2-2, предназначенного для ввода данных из внешних источников.
	* Указаны параметры для интерфейса: колонки 2-3 как классификационные, 4-10 как описательные.
	* Исходный XLS-файл сохранен в формате XLSX для уменьшения размера.
	* Проведена предварительная обработка данных: пустые ячейки (пробелы) в столбце "Взыскивался ли залог через суд" заменены на значение "нет", чтобы отразить отсутствие взыскания как информацию, а не отсутствие данных. Нули и пробелы в числовых данных настроены для интерпретации как отсутствие данных при вводе в систему.
3. **Формализация Предметной Области:**
	* Система автоматически создает классификационные и описательные шкалы с градациями на основе введенных данных.
	* Исходные данные (текстовые и числовые) кодируются с использованием созданных шкал и справочников.
	* Формируется обучающая выборка, где реальные значения заменены кодами.
4. **Синтез и Верификация Моделей:**
	* Запущен процесс синтеза и верификации моделей (режим 3-5) с использованием ускоренного алгоритма.
	* Создаются и верифицируются различные модели: 3 статистические (абсолютных частот, относительных частот PRC2, Inf1/Inf3/Хи-квадрат) и 7 системно-когнитивных.
	* Процесс занял 15 секунд.
5. **Анализ Моделей и Результатов:**
	* Проанализированы модели абсолютных и условных/безусловных относительных частот (PRC2).
	* Объяснены системно-когнитивные модели и интегральные критерии: "Сумма знаний" (скалярное произведение векторов объекта и класса) и "Резонанс знаний" (со стандартизацией векторов).
	* Оценена достоверность моделей с использованием F-меры Ван Ризбергена и ее нечетких/мультиклассовых обобщений, а также инвариантного критерия. Показано, что системно-когнитивные модели обладают большей достоверностью.
	* Проанализированы частотные распределения истинных и ложных решений, демонстрирующие рост доли истинных решений с увеличением уровня сходства.
6. **Прогнозирование и Исследование:**
	* Выбрана наиболее достоверная модель (Хи-квадрат с интегральным критерием "Сумма знаний").
	* Запущено решение задачи прогнозирования для всех 111 клиентов для определения вероятности возврата/невозврата кредита и ожидаемой прибыли.
	* Проанализированы результаты прогноза (степень сходства с классами) для отдельных клиентов.
	* Исследовано влияние отдельных факторов (признаков) на результат с помощью когнитивных функций и диаграмм. Определены наиболее и наименее значимые признаки и шкалы.
	* Проведен кластерный анализ классов и факторов для выявления их взаимосвязей.
7. **Сохранение и Заключение:**
	* Созданное приложение (модель и данные) сохранено в облаке Эйдос для дальнейшего использования и доступа.
	* Подчеркнута возможность адаптации и локализации модели путем добавления новых данных и удаления старых, что позволяет отслеживать динамику предметной области и применять модель в разных условиях.

**Детальная расшифровка текста:**

**1. Подготовка и Настройка**

Началось. Для чего?

Значит, у нас две классификационных шкалы будет. Две классификационных шкалы. Вторая, третья. А с четвертой по десятую описательные шкалы.

Посмотрите, сейчас мы запустим систему Эйдос. Удалим все приложения.

Так, что-то я не понял. Там какое-то мне сообщение, что ли? Сейчас я посмотрю. То есть вы начали запись. Вот.

**2. Ввод и Подготовка Данных**

* **Импорт из Excel и Параметры Интерфейса**
И я в этот чат буду отправлять скриншоты. И создаем приложение скоринговое по оценке риска невозврата кредита физическим лицам.

Для ввода информации из экселевской таблицы вот этой, мы используем программный интерфейс 2-3-2-2, предназначенный для ввода данных из экселевских файлов, в частности. 2-3-2-2. С параметрами, которые я сейчас задам. Значит, с четвертого по десятое. Сейчас уже, честно сказать, забыл, какие там параметры. Значит, вторая, третья колонка классификационная, с четвертой по десятую описательная. Как выяснилось, не забыл я. Вот. И вот с такими параметрами мы вводим данные.

* **Сохранение в XLSX и Обработка Пробелов**
Да, сейчас я еще вот что хочу сделать. Сейчас я его запишу как XLS файл, а потом открою в новой версии Экселя и уже сохраню его как XLSX. Потому что в новых версиях Экселя сокращается размер файла примерно в два раза. Поэтому есть прямой смысл это делать. Вот. И описание сейчас мы сюда скопируем.

Для чего? Для того, что когда мы сделаем приложение, мы, может быть, его запишем в облаке, это приложение. Вот. И с этим вот, с этими параметрами. Да, теперь новый Excel с этими параметрами будем вводить данные. Причем нули и пробелы будем считать отсутствием данных, а у нас их довольно много здесь пробелов, по крайней мере, вот.

Ну, правда, я всегда рекомендую, здесь вот у нас есть пробелы. Давайте мы эти пробелы заменим на "нет". Не знаю, получится или нет, посмотрим сейчас. Не получилось. Тогда просто наберем "нет".

Чтобы было не отсутствие информации, а была информация, что не взыскивали через суд. Это ж тоже информация, правильно? Взыскивали - информация, не взыскивали - тоже информация.

Мы видим, что у нас в исходных данных есть и данные текстовые, и числовые.

Пишем "нет". И только одну клеточку берем. Где тут еще? Кто не очень думая, сделаем.

Данные реальные, ребята.

Думаю, что лучше тонкие линии сделать, не так ужасно выглядит.

**3. Формализация Предметной Области**

* **Создание Шкал и Кодирование Данных**
Вот. Значит, дальше что мы делаем? Дальше мы входим в систему. Да, этот экранный интерфейс режим 2-3 2-3-2-2, который представляет собой программный интерфейс ввода данных из внешних источников данных, в частности экселевских таблиц. Данные могут быть текстовыми и числовыми.

Вот, вводим эти данные.

* **Формирование Обучающей Выборки**
Ну, в данном случае у нас здесь задано по три градации шкал числовых и классификационных, и описательных. Ну я думаю, что можно и побольше было бы сделать, 10, например. Вот, ну а здесь пять. Тогда мы пересчитаем это.

И выходим на создание модели. Создаются автоматически классификационные и описательные шкалы и градации. И потом они используются для кодирования исходных данных вот этих. И получается обучающая выборка. Это происходит быстро, даже на очень больших объемах данных. Это является важным плюсом системы Эйдос.

* **Анализ Классификационных Шкал**
Смотрим, ребята, какие у нас классификационные шкалы. Прибыль фирмы и возврат-невозврат. Это классификационные шкалы. Мы задали 10 диапазонов равной величины от минимальной до максимальной прибыли, которая была получена вот с этими клиентами. Возврат-невозврат.
* **Анализ Описательных Шкал**
Что-то вас совсем мало там осталось на собрании. Смотрим описательные шкалы. То есть что повлияло на прибыль и на возврат-невозврат? Залог, вид залога, пол, возраст, годовой доход, количество населения, взыскивался ли залог через суд.

Вот, здесь мы задали пять интервальных значений для числовых шкал. А текстовые шкалы, они так и есть. Уникальные текстовые значения, которые там были, они есть.

* **Просмотр Обучающей Выборки**
Смотрим обучающую выборку. Клиенты, здесь фамилии заменены, естественно, их реальные данные. И к каким классам они относятся, и какие у них значения свойств у этих клиентов. То есть исходные данные, которые мы видели, были закодированы системой, программным интерфейсом. И теперь мы видим вместо чисел и текстов, мы видим коды этих значений соответствующих свойств в соответствии со справочниками, которые я и в сам экселевский файл тоже исходных данных скопирую, фрагменты его, по крайней мере. Тоже скопируем в чат. Таким образом у нас произведена формализация предметной области.

**4. Синтез и Верификация Моделей**

* **Обзор Процесса АСКАнализа**
То есть мы выполняем вот эти этапы автоматизированного системно-когнитивного анализа, которые здесь у нас приведены в режиме 6-4. Ну сейчас я в чат пошлю из MS Visio изображение качественное, векторное. А в самой системе Эйдос тоже есть режим 6-4, где мы видим это изображение.

Значит, мы выполнили этап формализации предметной области. Теперь у нас следующий этап системно-когнитивного анализа - это синтез-верификация моделей. На этом этапе мы создаем три модели статистических и семь системно-когнитивных. Потом выбираем наиболее достоверную модель и в ней решаем задачи распознавания, прогнозирования, в данном случае прогнозирования, принятия решений и задачи исследования моделируемой предметной области. Режим 6-4.

* **Запуск Синтеза Моделей**
Для этого мы осуществляем сначала синтез моделей. То есть мы формализацию предметной области провели. Теперь осуществляем синтез моделей. Делается в режиме 3-5. Мы используем алгоритм, ускоренный алгоритм, в котором несколько упрощенно рассчитывается сходство объектов с классами при распознавании по интегральному критерию резонанс знаний, а сумма знаний точно рассчитывается, так же как и в классическом варианте.
* **Процесс Синтеза и Верификации**
Вот. Сейчас созданы все модели, а потом все эти модели по очереди верифицированы, то есть проверены на достоверность отнесения объектов к классам, к которым они относятся. Процесс занял 15 секунд у нас. Синтеза и верификации моделей.

**5. Анализ Моделей и Результатов**

* **Статистические Модели (Частоты)**
Смотрим на эти модели теперь. Модель абсолютных частот. Здесь мы видим, какое количество раз каждое значение свойства клиента встретилось в различных э-э при различных ситуациях результирующих: при различной прибыли и при различном возврате-невозврате кредита.

Так, что-то здесь интересное у нас. Возврат срок да, нет. А почему у нас еще есть да, нет? А почему их так много? Надо же, честно говоря, я не знаю. Почему так много у нас колоночек? Да, нет, и еще пустые какие-то. Надо будет посмотреть. Сейчас посмотрим.

Возврат срок. Возврат срок да и нет. Что-то неверно. Нам нужно по-другому создать модель. Сейчас мы ее удалим. 2-3-2-2. Здесь я понял, в чем дело. Здесь вот было э-э не выделять уникальных значений, не сортировать. А нам нужно их выделять и сортировать. Выделять уникальные значения и сортировать. И потом уже отменить эти параметры. И еще раз создадим. Модель. И сейчас мы увидим ситуацию другую. Значит, у нас да, нет, возврат. И здесь тоже у нас будет уникальных значений не будет итераций. Всё, понятно. Вот, теперь мы создаем модели.

Смотрим на эти модели. Модель Abs. Теперь мы здесь уже не увидим большого количества этих да, нет. Всё как надо получилось.

Мы смотрим, что у нас разные количество примеров по разным классам. По минимальной прибыли 21, по низкой прибыли 25 и так далее. И мы видим, что по разным прибылям у нас разное число примеров. И также разное число примеров по возврату в срок и невозврату. 60-51. Это означает, что мы э-э не не должны использовать эту матрицу абсолютных частот для принятия решений, а должны перейти к относительным величинам, что и сделано в модели PRC2. В этой модели мы видим, сколько процентов от числа наблюдений вот такой прибыли, сколько процентов клиентов, которые привели к такой прибыли, значит, э-э имели те или иные значения свойств. Это уже сравнивать более корректно эти числа, потому что здесь уже не играет роли количество клиентов по каждой категории. Но дело это неблагодарное, это сравнение. То есть это трудоемкая задача, сложная по объему. Это модель Inf1, это модель Inf3, квадрат.

* **Системно-Когнитивные Модели и Интегральные Критерии**
Возникает вопрос: что это за модели? Естественно, да? Значит, я вам основные модели показал. И сейчас объясню, что это за модели. Модель абсолютных частот - это сколько раз наблюдалось то или иное значение свойства клиента при том или ином результате кредита: возврат-невозврат, прибыль. Посчитано фактическое число встреч каждого значения свойства при каждом варианте исхода кредитования. Эта матрица стандартная, рассчитывается практически во всех системах статистических и в интеллектуальных тоже что-то похожее часто делается. Но, как я вам сказал, сравнивать абсолютные частоты не всегда уместно, хотя это можно делать корректно, квадрат на этом основан. Вот, а более интересно сравнивать условные, безусловные относительные частоты, которые стремятся к вероятностям при увеличении объемов выборки. Поэтому рассчитывается эта матрица в системе Эйдос. Таких систем, где она рассчитывается, уже поменьше. И потом на основе этой матрицы условных, безусловных относительных частот рассчитывается матрица системно-когнитивных моделей, что является специфической особенностью системы Эйдос. Это делается с помощью вот этих вот преобразований на основе матриц абсолютных частот и матрицы условных и безусловных процентных распределений. Если мы сравниваем два числа, то это можно делать путём вычитания и деления. Если мы путём вычитания сравниваем абсолютные частоты, фактические и теоретические, тогда получается квадрат. Если мы путём деления сравниваем фактические и теоретические частоты по формуле, которая здесь вот даны в начале по Пирсону, то тогда получается мера целесообразности информации Кракевича или коэффициент возврата инвестиций ROI (Return on Investment). И если мы сравниваем относительные частоты путём деления, у нас получается мера Кракевича или мера ROI, коэффициент возврата инвестиций. Если же мы сравниваем их путём вычитания, у нас получается коэффициент взаимосвязи. То есть я вам привел все модели, которые в настоящее время есть в системе Эйдос, как они рассчитываются. В результате мы получаем системно-когнитивные модели, которые и используются в системе Эйдос для решения всех задач, там решаются, которые я уже перечислил: группа задач, задача идентификации, прогнозирования, принятия решений и исследования моделируемой предметной области.
* **Оценка Достоверности Моделей**
Теперь смотрим, какова же достоверность моделей, которые мы создали. Мы видим, что достоверность рассчитывается по разным критериям в системе Эйдос. Основными являются F-мера Ван Ризбергена, классический критерий достоверности моделей, широко используемый. Но этот критерий обладает безусловными достоинствами, а также некоторыми недостатками. Недостаток у него заключается в том этого критерия, что, ну, недостатки точнее, что он является моноклассовым, чётким и зависит от объёма выборки, распознаваемой выборки. Здесь в хелпах коротко об этом говорится, который я привёл. Я подробно не буду это освещать, потому что я на лабораторных работах подробнейшим образом это описывал. Вот, но могу вам сказать, что по F-мере Ван Ризбергена 0,855 достоверность - это неплохая достоверность. И по критерию нечёткое мультиклассовое обобщение F-меры Ван Ризбергена уже 925. Это значительно выше, потому что более справедливые оценки используются. Система Эйдос сама имеет внутренний критерий достоверности, и если к ложным решениям суммировать, сумматором ложных решений положительных и отрицательных суммировать не единички, как предлагает Ван Ризберген, а оценку системы Эйдос степени принадлежности объекта к классам, то тогда мы видим закономерность такую, что при ошибочных решениях уровень сходства ниже, чем при истинных. Поэтому получается, что достоверность модели повышается. И предложено мною также обобщение критерия Ван Ризбергена, не зависящее от объёма выборки, распознаваемой, инвариантно относительно объёма выборки.
* **Частотные Распределения Решений**
Посмотрим на частотные распределения истинных и ложных решений. Сейчас я запишу название приложения, чтобы во всех выходных формах было это название. Название скоринговая система на основе АСК-анализа. Вот. И, значит, мы в результате видим наиболее достоверные модели, которой является модель квадрат с интегральным критерием сумма знаний. Мы видим, что доля истинных решений, которые здесь изображены красным и коричневым, значительно выше, чем доля ложных решений, которые изображены, количество которых изображено синим цветом. И, значит, видна явная закономерность, что при увеличении уровня сходства доля истинных решений закономерно растет, ну, положительных решений.

**6. Прогнозирование и Исследование**

* **Запуск Прогнозирования**
То есть мы видим из этих графиков, что доля истинных решений и положительных, и отрицательных выше, чем доля ложных решений. И доля истинных решений зависит пропорционально, по сути дела, уровню сходства. То есть уровень сходства можно рассматривать как адекватную меру степени истинности решения и степени риска, если решение, если класс соответствует невозврату. Ну и дальше мы должны решать задачу прогнозирования. Эту задачу мы будем решать в наиболее достоверной модели. Это в соответствии с нашим планом вот этим, который я привёл в виде схемы. То есть мы сделали модель текущей, в которой будем решать эту задачу. И потом запускаем само решение этой задачи ускоренным алгоритмом на центральном процессоре, которое я сделал вчера утром. Ну еще немножко не доделал. Здесь у нас хорошо всё получается. Вот идет прогноз для всех 111 клиентов возврата-невозврата кредита и прибыли для фирмы за счет взаимодействия с данным клиентом. За 2 секунды мы это получили результат для 111 человек.
* **Анализ Результатов Прогноза**
И посмотрим на этот результат. 4 1 3 1. Значит, мы получили для каждого клиента степень его сходства с классами. Для клиента один прибыль минимальная, один из десяти, не вернул в срок. И вот так вот для разных клиентов мы видим результаты. Вот для клиента 22 мы видим, что не вернул в срок и прибыль низкая, если из десяти градаций, то это первая градация.
* **Когнитивные Функции и Значимость Признаков**
Значит, мы сейчас я вам объясню, каким образом система определяет степень сходства клиента в данном случае, объекта распознавания с классом. Она это делает с помощью двух интегральных критериев. Что собой представляет интегральный критерий? Это способ приписать одно число объекту распознаваемому или клиенту в данном случае, которое характеризует степень его сходства с каждым из классов, которые у нас сформированы в модели. В Эйдос используются два интегральных критерия. Один интегральный критерий называется сумма знаний. У нас в матрицах модели содержатся знания. Понятие данных, информации, знания. Сейчас я вам, ребята, данные, информация, знания. Сейчас я попробую вам кинуть отношение, картинку соотношений данных, информации, знания. Значит, данные - это неосмысленная информация, сырая информация. Когда мы данные осмысливаем, то получается уже информация. То есть информация - это осмысленные данные. Смысл данных в том, что мы знаем, какие причинно-следственные связи в той предметной области, которая описывается этими данными. То есть когда мы узнаём, какие причинно-следственные связи, силу их и направление, то мы, по сути дела, преобразуем данные в информацию. В системе Эйдос эта операция производится. Она подготавливается на этапе формализации предметной области, когда мы разрабатываем справочники прошлых и будущих событий. И потом используем их для кодирования исходных данных. А потом формируем модели вот эти, которые я показывал, статистические и системно-когнитивные, которые уже и отражают силу и направление причинно-следственных связей. То есть как влияет каждое значение фактора на принадлежность объекта моделирования или переход объекта моделирования, в качестве которого у нас клиент выступает, в состояние, соответствующее классам. А состояние у нас - это градации по прибыли, которую получает фирма, 10 градаций, и диапазоны числовые, и возврат-невозврат кредитов в срок. Если мы эту информацию используем для принятия решений, ну, скажем, о выдаче или невыдаче кредита, тогда эта информация превращается в знания. Поэтому, то есть знание - это информация, полезная для достижения цели, повышения прибыли, к примеру, и уменьшения случаев невозврата кредитов в срок. Взять эту задачу. Таким образом, мы должны определить, какое суммарное количество информации содержится в признаках заёмщика о том, что он вернёт или не вернёт кредит и о том, какая будет получена прибыль. То есть мы должны посчитать суммарное количество информации в его признаках о каждом из этих исходов кредитования, не только о благоприятных или неблагоприятных, а о всех. Это делается очень просто. Просто мы суммируем количество информации, которое есть в признаках этого конкретного заёмщика, о принадлежности его к тем или иным классам. В разных моделях мы это делаем. Ну, и по признакам, которые есть у заёмщика, мы суммируем информацию, а по признаку, которого у него нет, не суммируем. Математически это можно выразить следующим образом. Если мы опишем заёмщика некоторым вектором, координаты которого будут равны нулю, если у него нет признака соответствующего, а каждый, каждая координата этого вектора соответствует какому-то признаку. И равно единице, этот вектор равен единице, если этот признак у него есть у заёмщика. Ну это я немножко упрощённо рассказываю. То тогда сумма произведений элементов матрицы модели, если у неё и-тая строчка - это соответствует какому-то значению свойства, а G - классу. Вот такая сумма произведений координат этих векторов, описывающих объект идентифицируемый, распознаваемый, то есть заёмщика в данном случае, и вектора класса, координат вектора класса. Вот, то тогда мы получим суммарное количество информации в признаках объекта о принадлежности его к данному классу. Эта форма, она часто встречается математическая в самых разных областях. И даже вот в теории позиционных систем счисления. То есть очень широко распространена подобная формула, имеющая разный смысл в разных областях. Но если подумать, то, в общем-то, очень интересные напрашиваются ассоциации, аналогии, содержательные, осмысленные. Но известно, что такая форма - это произведение, скалярное произведение двух векторов в координатной форме. Это интересная математическая особенность этого интегрального критерия, благодаря которой он не является мерой расстояния. А раз он не является мерой расстояния, ну так наподобие, скажем, эвклидова расстояния, то значит, для него никакого отношения к нему не имеют те свойства пространства, которые являются обязательными для использования тех или иных мер расстояния. Ну, скажем, вот эвклидово расстояние - это мера, которая корректна в ортонормированных пространствах, в которых оси взаимно перпендикулярны, расстояние Пифагора. Наша же мера сходства двух векторов, она корректна и для неортонормированных пространств. Она является, эта формула является формулой, определяя, по сути дела, определением белого шума. Если взять случайный сигнал, то любые его два фрагмента одинаковой длины, если их рассматривать как координаты векторов и посчитать скалярное произведение, то этот, это скалярное произведение будет стремиться к нулю, а угол между этими векторами будет стремиться к 90°, при увеличении числа координат, при увеличении длины векторов. Это есть определение шума, собственно говоря, белого. Это что значит? Что если у нас в исходных данных есть шум, а он там всегда есть, как мы понимаем, это принципиально неустранимая компонента исходных данных. То есть в исходных данных всегда есть и полезный сигнал, и шум. То, значит, вот такой интегральный критерий увеличивает отношение сигнал/шум, подавляя шум и суммируя в матрице абсолютных частот частоты, мы получаем увеличение сигнала, амплитуды сигнала по отношению к шуму, потому что шум при суммировании стремится к нулю, это его определение, а при, а значит, полезный сигнал суммируется просто как умножается на число наблюдений. Этот интегральный критерий имеет сходство по смыслу с функцией принадлежности элемент к множеству, которое применяется в нечёткой логике, в частности вот в её варианте, который разработан Лотфи Заде. Однако в нечёткой логике этот интегральный критерий, функция принадлежности точнее, просто несколько вариантов им предложено самим разработчиком нечёткой логики и другими учёными, которые в этой области работали. И просто он выбирается вид функции принадлежности из нескольких вариантов. Выбирается самим пользователем. То в системе Эйдос значение интегрального критерия рассчитывается на основе моделей, полученных непосредственно на основе эмпирических данных. То есть, если в нечёткой логике мы его выбираем априорно просто на основе каких-то соображений, то здесь он рассчитывается на основе эмпирических данных. Я считаю, что это одна из проблем нечёткой логики - выбор функции принадлежности, выбор вида её и расчёт её значения. Здесь эта проблема решается у нас. Поскольку мы видели, что доля истинных решений растёт пропорционально уровню сходства, то это означает, что уровень сходства является адекватной мерой степени истинности решения или риска ошибки. То есть в системе Эйдос есть внутренний критерий аудиторский, можно сказать, самооценки степени истинности тех решений, которые она предъявляет, предлагает. И очень важно отметить, что эти коэффициенты, они сходны по своей форме с коэффициентами разложения в ряд. По этому поводу я написал книжку недавно, издал в 2020 году, которую вам очень рекомендую ознакомиться с этой книжкой, с монографией. Сейчас я попробую её выцарапать ссылочку на неё. Это сделаю со своего сайта. Возьму эту ссылку. Пункт два. Скачать, запустить. Вот, сценарный. Вот, сценарный метод АСК-анализа. В этой книжке подробно описывается математика, лежащая в основе этого метода, который применяется в системно-когнитивном анализе и в системе Эйдос. И применение, примеры применения.
* **Резонанс Знаний и Стандартизация**
Второй интегральный критерий очень схож с первым, который я описал. Тоже является скалярным произведением двух векторов: объекта и класса. Но вектора стандартизированы тем способом, который применяется в статистике. То есть из каждой координаты вычитается среднее и делится на среднеквадратичное отклонение. Я бы сказал ещё вот что, что есть много разных способов стандартизации. Я бы сказал даже, что их бесконечное количество. Но один из таких наиболее известных, кроме вот того, который применяется в статистике, само значение координаты, из него вычитается среднее по вектору и делится на среднеквадратичное отклонение. Ещё используется такой способ стандартизации широко, который связан с линейным, с линейной интерполяцией, то есть линейными сплайнами, то есть заменой фрагмента функции прямыми линиями, грубо говоря. Вот. То есть тоже значение координаты вычитается минимальное и делится на максимальное минус минимальное. Вот. То есть таких интегральных критериев можно придумать неограниченное количество: аддитивных, мультипликативных. Если мы складываем, вот там я вам показывал, да, формула сложения. Вот. Сложение координат, сумма произведений координат. То можно, конечно, было бы их не складывать, а перемножать вот эти произведения координат. Тогда был бы мультипликативный критерий, а не аддитивный. И можно и другие формы придумать общего вида. Но могу вам сказать, что в системе Эйдос только два используются аддитивных интегральных критерия, которые я вам уже показал: сумма знаний и резонанс знаний. Значит, резонанс знаний - это интегральный критерий, который рассчитывается по всей матрице. А сумма - только по тем признакам, которые есть у объекта распознаваемого. Поэтому с интегральным критерием сумма знаний можно очень быстро посчитать результат, степень сходства объекта с классом. Только взяв только суммы информации или знания вот из матрицы модели по тем признакам, которые есть. А вот здесь мы с интегральным критерием резонанс знаний, мы должны сумму взять по всем признакам, которые есть не у объекта, а в модели. Поэтому он может расчёт занимать, может, значительно больше времени. Ну можно упрощённо это сделать. Упрощённо это сделано в режиме ускоренного расчёта. Там сумма знаний адекватно рассчитывается, а резонанс знаний, ну, правдоподобно рассчитывается, я бы сказал так, оценка даётся его ускоренная. А по полному перечню, если это делать в классическом варианте, то это может занимать довольно заметное время, если много признаков модели, если она большой размерности. Вот. Ну и эти два интегральных критерия используются, а вот другие способы нормировки векторов не используются в системе Эйдос. Я только один способ использовал. Хотя я знаю, что их очень много.
* **Когнитивные Диаграммы и Нейронные Сети**
И вот так решается задача прогнозирования прибыли и самого факта возврата и невозврата. Теперь давайте посмотрим, что там у нас за модели сформировались. Значит, для этого мы используем режим 448. Смотрим, когда у нас получается минимальная прибыль. Она у нас получается тогда, когда залог взыскивается через суд. Вот. Количество населения в фактическом месте жительства. Ну тут вот, может быть, имеет смысл такую диаграмму посмотреть. Это когда низкая прибыль. Меньше 50 человек, то есть маленькие населённые пункты. Залог - участок, возраст клиента минимальный, залог, вид залога - недвижимость, доход клиента минимальный, возраст либо вообще минимальный, либо молодой. Ну вот и судите сами, давать ему кредит или нет таким заёмщикам. То есть мы вот такую информацию уже из модели получаем, на основе которой может приниматься решение разумное. Как влияет любой признак? Вот возраст большой, как влияет? Он влияет вот так, что возврат в срок будет, прибыль будет низкая или чуть выше средней. Вот так вот. В основном это влияет на то, что будет возврат в срок, наиболее сильно. То есть мы не просто получаем прогноз, вернёт, не вернёт и какая прибыль, а мы получаем еще модель, которую можно использовать для анализа. И, допустим, смотрим возврат в срок, нейрон в форме нейрона представление. Значит, здесь мы видим, как какие рецепторы активируют и тормозят этот нейрон. И видим все нейроны вместе в виде нейронной сети, модели F3. Здесь мы видим синими линиями обозначены тормозящие воздействия, красными - активирующие, толщина линии отражает силу воздействия. 2/3 модели мы видим в этом в этом одном слое этой сети.
* **Кластерный Анализ**
Смотрим на когнитивные функции. Значит, когнитивные функции отражают, какое количество информации содержится в тех или иных признаках объекта о том, что будут получены по результат, по следствие их влияния будут получены определённые результаты, соответствующие классам. Мы видим, значит, всю эту информацию. Как влияет вид залога, как влияют другие факторы. В любой модели можно посмотреть. Значит, вот если залог автомобиль, тогда прибыль низкая будет. Если в залог недвижимость, тогда, нет, извините, если залог недвижимость, тогда прибыль будет низкая. Если залог автомобиль, прибыль будет очень высокая фирмы. Вот. И так мы видим по всем сочетаниям шкал. Мы можем задать режим, когда все эти формы будут сейчас записаны в папочку. Я оттуда их возьму. Количество этих форм будет равно произведению числа классификационных и описательных шкал.

Можем посмотреть, когда возврат в срок происходит. Вот мы получаем, когда через суд нет, когда возраст пожилой, когда большой населённый пункт, средний доход, залог, когда квартира даётся, когда высокий доход или средний. И когда пол женский. Вот такие вот дела, ребята. То есть мы можем теперь определить, сколько у нас до конца занятия, скажите? 20 минут.

То есть мы можем оценить значимость вот этих вот свойств клиентов, рассортировать их в порядке убывания значимости, просуммировать значимость нарастающим итогом. У нас получится, ребята, смотрите, треть, 28% наиболее значимых признаков обеспечат 50% суммарной значимости, а половина наиболее значимых признаков обеспечит 72% суммарной значимости. То есть это значит, что у нас есть признаки более ценные и менее ценные для решения наших задач. А признаки являются градациями шкал, ребята, описательных. И мы можем оценить и значимость самих шкал, как среднее от значимости их градаций. И у нас есть соответствующие экселевские файлы, в которых мы видим эту информацию. Посмотрим на эти файлы. Значит, если мы рассортируем в порядке по времени, то мы увидим новые базы данных, созданные только что. Вот, допустим, видим экселевский файл F3. Значит, я могу вам сказать, что наиболее значимыми шкалами являются: взыскивается ли через суд, гарантия возврата, и какое население, какой годовой доход. Вот это является наиболее ценным. Вот значимость первой градации взыскивается через суд - 33-34%, то есть треть. А вот вид залога и пол клиента - это малозначимые свойства. Годовой доход наиболее значимый, если не считать население в пункте, в месте жительства. То есть если приезжают к вам из маленьких каких-то населённых пунктов и занимают деньги молодые люди мужского пола, то это наибольшее даёт информацию нам о том, вернут они или не вернут кредит. Как мы знаем, они его не вернут при таких параметрах. И то же самое мы можем сказать о степени детерминации классов. То есть у нас факторы, влияющие на то, что возникнут те или иные состояния, они либо более жёстко, либо не очень жёстко обуславливают эти состояния. Некоторые будущие состояния объекта моделирования жёстко обусловлены факторами, свойствами клиента, а некоторые нет. И мы можем посмотреть в экселевских формах, какие именно наиболее жёстко обусловлены состояния. Сейчас я вам покажу это, и на этом мы закончим занятие. Мы видим, что наиболее жёстко обусловлен возврат кредит в срок, возврат кредитов в срок и возврат, э-э, невозврат кредитов в срок. А также минимальная прибыль, ребята. Минимальная и низкая прибыль тоже жёстко обусловлены. То есть они легко устанавливаются. А вот высокая прибыль и средняя, и выше среднего, они э-э слабо детерминированы, то есть могут быть разные варианты. А вот очень низкая прибыль устанавливается практически без безошибочно. И сам возврат и невозврат кредитов в срок тоже устанавливается безошибочно, ребята.

Диаграмму я вам покажу, то есть как связаны классы друг с другом по системе их детерминации. И мы видим, что низкая прибыль, невозврат и очень низкая прибыль, они обусловлены одними и теми же значениями факторов. Мы также можем это увидеть в кластерном анализе. Проведём кластерный анализ. И сейчас я его формы тоже в чат помещу. Здесь мы видим, что очень низкая, невозврат кредита, очень низкая прибыль и прибыль, они в один кластер объединены, а другие варианты в другой кластер объединены. То есть у них совершенно различная система детерминации вот этими свойствами заёмщика, заёмщика. И сейчас я еще раз нарисую без перерасчёта для того, чтобы график межкластерных расстояний поместить в чат. И аналогичный кластерный анализ мы можем провести для значений факторов. То есть мы можем определить, какие свойства заёмщиков имеют сходный смысл. Здесь мы мало что увидим. Вот, ну правда, можно задать параметры, такие, при которых увидим. И более, ну, дело вкуса, конечно, но более такая форма интересная - это форма кластерного анализа значений свойств заёмщика, значений факторов, влияющих на прибыль и на возврат и невозврат кредита. Мы видим, какие сходные. Вот пол клиента мужской и очень низко, очень молодой, и взыскивается через суд, и население очень маленькое. Всё это вот образует один кластер, а другие факторы - другой. То есть мы уже поняли, какой именно смысл этого кластера. Сначала классы объединяются в кластеры на высоком уровне сходства, потом это становится сложным, не вообще невозможным, и тогда формируются полюса конструкта, то есть система противоположных кластеров.

**7. Сохранение и Заключение**

* **Сохранение Приложения в Облаке**
Вот, ребята. А теперь мы сейчас что сделаем? Сейчас мы зайдем в режим 1-3 и запишем приложение в облако, то, которое у нас сейчас создано. Вот эти файлы будут записаны в облако. Их размер суммарный - 10 МБ. Начинается процесс записи. Значит, добавляем, этот процесс описан в помощи хорошо. Добавляем новую строчку для приложения, копируем информацию с предыдущей строчки. Можно задать любую строчку, из которой мы копируем информацию. И сохраняем приложение в облаке. Сохранение идёт файлов сейчас по FTP протоколу, запись в Эйдос облако этого приложения. После этого, да, название, естественно, берётся из названия того, которое мы ввели в режиме 1-3. Естественно, вы теперь можете это приложение скачать, установить у себя на компьютере и исследовать. Описание немножко не соответствует, потому что та статья, я сейчас немножко исходные данные скорректировал, и когда вводил информацию в систему, то тоже не обращал внимания на то, как это делается в статье. То есть я другое количество градаций задал числовых шкал. И, как вы помните, задал не только да, но и нет параметр. Что более правильно. Файлы, которые туда записываются в облако, не могут быть больше 10 МБ по размеру, ребята. Вернее так, они могут быть больше 10 МБ, они запишутся, но они сразу сотрутся хостингом самим, моментально. Всё, процесс завершён. Для его закрытия надо выйти из этого режима. Записывается каталог приложений, после чего у нас в облаке появилось приложение, ребята, которое только что я вам рассказывал. Вот оно, 298. Если по нему кликнуть два раза по этому приложению, то откроется статья, потому что туда она записана в виде PDF-файла с описанием приложения. И там же есть и вордовский файл. Вот такая вот, ребята, картина.
* **Вопросы и Ответы, Адаптация и Локализация**
Какие у вас есть вопросы? Пожалуйста, задавайте вопросы. Что-то как-то вы не задаёте вопросов никаких. Вот. Ну тогда ещё вам… Что-что? По этому нет. Вопросов нет? Ну я показал вам, вы же экономисты, показал вам, как разрабатывать с помощью системы Эйдос риэлторские приложения, как разрабатывать оценку риска банкротства, модели оценки риска банкротства, как разрабатывать модели риска кредитования и невозврата кредита, вот сейчас вам показал. То есть я вам показал ряд экономических приложений, которые можно, интересных, практически значимых приложений, которые можно, для создания которых можно использовать систему Эйдос. И создание, и применение. И что ещё интересно, ребята, я хотел бы вам дополнить, ещё, что эти приложения можно эксплуатировать в адаптивном режиме. То есть можно добавлять туда информацию по новым клиентам в базу данных и пересоздавать модели. В результате у нас модель будет отслеживать динамику предметной области. Старые данные можно удалять из модели, чтобы они нам не портили картину. Мы можем использовать методику в адаптивном режиме эксплуатировать. То есть можем добавлять в базу исходных данных, добавлять информацию о новых клиентах. Слышно, нет? Слышно. Вот. И старых удалять. В результате у нас получится, что данные будут постоянно актуальными, и модель будет постоянно отражать динамику предметной области. Также вот, возможно, эту модель захотят использовать в других регионах, может, даже в странах. А там зависимости между свойствами клиентов, возвратом-невозвратом кредита и прибылью могут быть другими, чем у нас. Тогда то же самое, если делать, эксплуатировать приложение, начинать его с того, что сейчас я вам показал, а потом добавлять туда данные уже их региона и новые, а старых потихонечку удалять. Ну, оставить там, скажем, примеров 200 клиентов или 300, и хватит. Новых добавлять, старых удалять. Вот. То через некоторое время эта модель адаптируется, то есть локализуется, то есть она будет отражать уже закономерности не Краснодарского края, города Краснодара, а того региона, где мы будем её эксплуатировать. Я скажу так, что достоверность её повысится в результате. Конечно, наши закономерности, они где-то и в других местах тоже действуют, но с несколько с другой силой. И эта вот сила будет корректироваться, сила и направление влияния этих вот свойств будет корректироваться при таком подходе. То есть я вам хочу сказать, что система Эйдос обеспечивает не только создание приложений, но и их эксплуатацию, потому что она является инструментом их создания. А и также эксплуатацию в адаптивном режиме, потому что она является инструментом их создания и адаптации. Это очень важные моменты. Это значит, что то, что я сейчас вам рассказываю, не только для вас, может быть, интересно, а допустим, для тех, кто живёт в Перми или в Петербурге. Мы также можем и локализовать и адаптировать этим путём. То есть адаптация - это учёт динамики предметной области во времени, а локализация - учёт особенностей местности, где происходит эксплуатация.
* **Заключение и Дополнительные Ресурсы**
Ну, надеюсь, вам понравилось само, сам способ разработки скорингового приложения в системе Эйдос, возможность его эксплуатации в системе Эйдос, возможность его эксплуатации в адаптивном режиме и адаптации и локализации его для других мест применения, не тех, где оно было создано, разработано. Что и обеспечивает его ценность для тех, кто, может быть, им воспользуется, посмотрев облачные Эйдос-приложения. Это могут сделать люди во всём мире, как вы знаете. Вот. И они этим, это и делают. Сейчас я вам покажу. Да, я знаю, сейчас. Вот. Всё, ребята, на этом у нас конец занятия. До свидания. Если будут, если будут вопросы, можете задавать эти вопросы и в самом Тимсе, и в чате, или, то есть не тогда, когда занятия, а вообще, в принципе. Вот. Пожалуйста, пожалуйста, я на них отвечу. И вот, ребята, как пользоваться системой Эйдос в мире? Многие пользуются. И вам советую тоже. Всё, самого хорошего. До свидания. До свидания. До свидания. До свидания.