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

***Российская Федерация***

**52 Интеллектуальные технологии и представление знаний. Лабораторная 3. вопрос достоверности моделей 2020-11-26**

## Лекция: Анализ моделей данных, частоты и оценка достоверности

### Резюме лекции

**1. Введение и организационные моменты:**

* Лекция проводится 26 ноября 2020 года, первая пара (8:00 - 9:30).
* Тема: Лабораторная работа №3 по дисциплине "Интеллектуальные технологии и представление знаний".
* План: Рассмотрение моделей частот и понятия достоверности моделей.
* Лектор: Профессор Луценко Евгений Вениаминович.

**2. Анализ моделей данных: от абсолютных к относительным частотам:**

* **Матрица абсолютных частот (сопряженности, корреляционная):** Представлена как основа для анализа. Столбцы - обобщающие категории (классы), строки - значения свойств (признаки).
* **Проблема абсолютных частот:** Некорректно делать выводы напрямую из матрицы абсолютных частот, так как количество объектов в разных категориях (классах) различно. Это искажает сравнение характерности признаков для разных классов.
* **Решение - относительные частоты:** Необходимо перейти к относительным частотам (например, процентам) для корректного сравнения.
  + **Условные относительные частоты:** Частота признака внутри конкретного класса.
  + **Безусловные относительные частоты:** Частота признака по всей выборке (средняя).
* **Сравнение частот:** Сравнивая условные частоты с безусловными и между собой, можно делать обоснованные выводы о характерности признака для класса.

**3. Проблема анализа больших матриц и автоматизированные методы:**

* **Сложность ручного анализа:** Визуальный анализ и сравнение частот в больших матрицах (сотни признаков, десятки/сотни классов) неэффективны и трудоемки.
* **Автоматизация сравнения:** Необходимы формальные меры для автоматического сравнения частот и выявления закономерностей. Предложены три подхода:
  + **Мера Пирсона (Хи-квадрат):** Разница между фактической и теоретической частотой признака в классе. Считается наиболее достоверной и четкой.
  + **Мера Харкевича (Количество информации):** Логарифм отношения условной относительной частоты к безусловной. Отражает количество информации, которое дает признак о принадлежности объекта к классу.
  + **Коэффициент возврата инвестиций (ROI):** Отношение частот минус единица. Аналогичен подходу Харкевича без логарифмирования.
* **Интерпретация мер:** Положительные значения указывают на прямую связь (признак характерен), отрицательные - на обратную (признак препятствует принадлежности), нулевые - на отсутствие связи.

**4. Достоверность моделей:**

* **Необходимость проверки:** Прежде чем использовать модель для решения задач (идентификация, прогнозирование, принятие решений, исследование), необходимо убедиться в ее достоверности.
* **Достоверность vs. Непроверенность:** Модель, достоверность которой не изучалась, может быть как достоверной, так и недостоверной. Использование непроверенной модели крайне рискованно.
* **Последствия использования недостоверной модели:** Ошибки в диагностике, неверные прогнозы, убыточные решения, ложные научные выводы. Приведены примеры из медицины, экономики, аналогия с непроверенными социальными экспериментами (Октябрьская революция) и сравнение применимости моделей (АТС vs Супермаркет).
* **Затраты на точность:** Повышение точности (достоверности) модели требует увеличения объема выборки, что связано с затратами (время, деньги). Необходимо находить баланс между требуемой точностью и затратами.
* **Оценка достоверности:** Простейший способ – проверка на ретроспективных данных с использованием задач идентификации (распознавания). Сравниваются результаты работы модели с известными фактическими данными.

**5. Вывод:**

* Анализ данных начинается с построения моделей частот, но для корректных выводов необходим переход к относительным частотам и автоматизированным мерам сравнения.
* Ключевым этапом является оценка достоверности модели перед ее практическим применением, чтобы избежать серьезных ошибок и неэффективных решений.

### Детальная расшифровка текста

**1. Введение и организационные моменты**

* **(0:02)** [звук вдоха]
* **(0:03)** [звук глубокого вдоха и питья воды]
* **(0:15)** Что?
* **(0:19)** Начнём занятие.
* **(0:21)** Что-то как-то вас маловато, по-моему.
* **(0:24)** [тишина, фоновый шум]
* **(0:34)** Ну что, значит, с чего мы начнём? С того, что
* **(0:38)** я скажу, какое сегодня число, дата и всё прочее там.
* **(0:42)** Сегодня 26 ноября 2020 года.
* **(0:47)** Первая пара.
* **(0:49)** 8:00.
* **(0:51)** 9:30.
* **(0:55)** По
* **(0:57)** дисциплине.
* **(0:59)** Лабораторная работа номер три по дисциплине
* **(1:03)** интеллектуальные технологии
* **(1:06)** и представление знаний.
* **(1:09)** И у нас
* **(1:11)** по плану
* **(1:13)** рассмотрение моделей
* **(1:15)** и потом рассмотрение понятия достоверность
* **(1:19)** модели.
* **(1:23)** Занятие ведёт профессор Луценко Евгений Венеаминович.
* **(1:28)** [тишина, фоновый шум]
* **(1:30)** [звук щелчка мыши или клавиатуры]
* **(1:36)** [тишина, фоновый шум]
* **(1:37)** Вот.
* **(1:40)** Здравствуйте, ребята. У меня микрофон выключен был. Здравствуйте.
* **(1:46)** Так, ну теперь…
* **(1:47)** Доброе утро, Евгений Венеминович. Всем доброе утро.
* **(1:50)** Доброе утро, Георгий Александрович.
* **(1:53)** [тишина, фоновый шум]
* **(2:02)** [неразборчивые звуки на фоне, возможно, разговор или шум]
* **(2:08)** Так, сейчас
* **(2:09)** сделаю экран.
* **(2:12)** [тишина, фоновый шум]
* **(2:17)** Вот.
* **(2:19)** Ну начнём мы с чего?
* **(2:23)** С того, что
* **(2:26)** установлю я лабораторную работу, на которой мы будем всё это рассматривать.
* **(2:32)** [тишина, фоновый шум]
* **(2:48)** [тишина, фоновый шум]
* **(2:50)** Уже я не буду это комментировать, потому что вы это всё знаете.
* **(2:55)** [тишина, фоновый шум]
* **(3:14)** [тишина, фоновый шум]
* **(3:18)** [тишина, фоновый шум, дыхание]
* **(3:32)** Кирилл, скоро у тебя будет там Mercedes фотография? А то что-то какая-то у тебя там… Что это у тебя за машина?
* **(3:42)** Гранта?
* **(3:45)** [тишина, фоновый шум]
* **(3:51)** Ну вот.

**2. Анализ моделей данных: от абсолютных к относительным частотам**

* **2.1. Матрица абсолютных частот (сопряженности, корреляционная):**
  + **(3:54)** Теперь смотрим модели.
  + **(3:57)** [тишина, фоновый шум]
  + **(4:08)** Мы остановились на чём? На том, что я показал модель абсолютных частот.
  + **(4:15)** Это ещё называется матрица сопряжённости.
  + **(4:19)** [звук уведомления]
  + **(4:24)** М-м-м. Ладно.
  + **(4:29)** Матрица абсолютных частот, матрица сопряжённости или корреляционная матрица.
  + **(4:34)** Как она построена? Значит, по колонкам у нас
  + **(4:39)** обобщающие категории, группы,
  + **(4:43)** которым принадлежат объекты выборки.
  + **(4:46)** А по строкам - значения свойств или значения факторов. Я их называю признаки, сокращённо.
  + **(4:53)** А колоночки называю классами.
  + **(4:56)** Просто, чтоб не говорить, что это там градация, классификационная шкалы там или там значение свойства. Просто я говорю: значение свойства - это признак,
  + **(5:05)** а градация классификационной шкалы - это класс.
  + **(5:10)** А градация описательной шкалы - это признак.
  + **(5:14)** [тишина, фоновый шум]
* **2.2. Проблема абсолютных частот:**
  + **(5:15)** Вот мы видим, что
  + **(5:18)** здесь у нас число встреч какого-то признака,
  + **(5:21)** каждого признака
  + **(5:23)** у объектов
  + **(5:25)** определённой категории.
  + **(5:28)** Я вам говорил, что на основе такой матрицы непосредственно, вот её, делать выводы не неосновательно. Почему? Нет возможности
  + **(5:37)** корректно делать выводы. Потому что по разным категориям разное число объектов. Вот, допустим, здесь мы видим,
  + **(5:43)** три раза встречается
  + **(5:46)** признак "размер под руку"
  + **(5:49)** и у средств связи, и у элементов компьютеров.
  + **(5:53)** И
  + **(5:56)** казалось бы, этот признак одинаково характерен для этих категорий объектов, но это не так, потому что
  + **(6:02)** средств связи три, а элементов компьютеров семь.
* **2.3. Решение - относительные частоты:**
  + **(6:06)** Поэтому мы должны перейти к относительным величинам.
  + **(6:11)** Ну, можно в процентах измерять.
  + **(6:15)** Это удобно.
  + **(6:18)** Вот. Если мы посмотрим на размер под руку,
  + **(6:22)** этот же самый признак, то у нас окажется, что 100% средств связи этим признаком обладают,
  + **(6:28)** и 42%, 8 десятых элементов компьютеров. То есть здесь уже ясно, что для
  + **(6:34)** средств связи он гораздо более характерен, чем для элементов компьютеров.
* **2.4. Сравнение частот:**
  + **(6:38)** И можем ещё посмотреть в среднем, насколько часто он встречается по всей выборке.
  + **(6:43)** То есть эти вот 100% и 42,8 - это условные относительные частоты.
  + **(6:51)** Условные, в смысле, они
  + **(6:53)** для объектов определённых категорий.
  + **(6:56)** При условии, что объект относится к определённой категории.
  + **(7:00)** А безусловная относительная частота - это по всей выборке, по всем категориям.
  + **(7:07)** Мы можем сравнить
  + **(7:08)** относительные, условные относительные частоты с безусловной и друг с другом. И сделать на основе этого вполне обоснованные выводы,
  + **(7:16)** что этот признак более характерен для средств связи,
  + **(7:19)** менее характерен для элементов компьютера, но всё-таки и для элементов компьютеров он более
  + **(7:26)** характерен, чем в среднем для всех объектов выборки.
  + **(7:31)** Ну то есть всё-таки это тоже это характерный признак и для элементов компьютера, просто менее характерный, чем для средств связи.
  + **(7:38)** Вот. То есть мы можем вот так сравнивать и делать выводы на основе этого.

**3. Проблема анализа больших матриц и автоматизированные методы**

* **3.1. Сложность ручного анализа:**
  + **(7:42)** И было бы всё ничего, если бы э-э матрицы эти могли быть очень… они могут быть очень большой размерности.
  + **(7:49)** Эти вот э-э
  + **(7:51)** признаки, их могут быть сотни тысяч.
  + **(7:54)** Ну, реально так, конечно, редко бывает. Реально бывает там 500, 700, там 1800. Ну, были случаи у меня, когда их было там и 11.000, например.
  + **(8:05)** А вот классов может быть в системе IDS вообще до 1.500 групп.
  + **(8:11)** Ну, обычно там их 20, 30, 40, там, ну 100, например, вот такое количество какое-то.
  + **(8:17)** То есть какие-то разумные
  + **(8:20)** количества.
  + **(8:22)** Ну, как-то раз было и 500, по-моему, 380 вот что-то такое, до 500.
  + **(8:27)** Так вот,
  + **(8:29)** если будет матрица 500, например, колоночек, здесь 14 колоночек, а представьте, что их 500, а строк, э-э, скажем, 11.000. Ну тогда понятно совершенно, что вот такое вот сравнение
  + **(8:42)** с помощью зрачков, как говорится, когда вот мы сидим и смотрим вот
  + **(8:48)** глазами просто на эту матрицу. И с помощью своего естественного интеллекта пытаемся всё это осмыслить, то это, конечно,
  + **(8:56)** дело бесперспективное, потому что
  + **(8:59)** очень большая размерность данных.
  + **(9:02)** Мы просто это не сможем прочитать. Если мы это распечатаем, будем читать, может быть, нам потребуется несколько лет, чтобы это прочитать, понимаете?
  + **(9:09)** Ну и даже, допустим, это и прочитаем. Ну и что? Мы же не сможем сделать выводы на основе этого.
  + **(9:14)** А если мы… а мы должны делать выводы на основе уже сравнений.
* **3.2. Автоматизация сравнения:**
  + **(9:20)** Поэтому мне пришла в голову такая мысль. Она пришла мне в семьдесят девятом году.
  + **(9:26)** Автоматизировать это сравнение.
  + **(9:30)** Как я это сделал?
  + **(9:33)** Я это сделал разными способами.
  + **(9:36)** Один способ общеизвестный - это
  + **(9:40)** хи-квадрат. Это разница между фактической частотой наблюдения этого признака в житом классе и теоретической частотой.
  + **(9:48)** Это э-э мера, разработанная Пирсоном,
  + **(9:52)** разработчиком вообще параметрической статистики всей.
  + **(9:56)** Прекрасная, прекрасно себя она показывает.
  + **(10:00)** Чаще всего эта модель, основанная на хи-квадрат, она самая достоверная и наиболее так чётко всё в ней
  + **(10:06)** просматривается, все закономерности.
  + **(10:10)** Но если мы хотим сравнить два числа,
  + **(10:13)** какое из них больше, какое меньше,
  + **(10:16)** то э-э, кроме вычетания,
  + **(10:19)** способа, который применил сам Пирсон,
  + **(10:21)** ещё есть возможность делить их друг на друга, да?
  + **(10:25)** Вот если мы сравниваем два числа 7 и 5, то какое число больше?
  + **(10:29)** Вычитаем из семи пять,
  + **(10:31)** получается больше нуля, значит семь больше.
  + **(10:35)** А если мы путём деления их сравниваем, тогда 7 делим на 5, получается величина больше единицы, значит семь больше,
  + **(10:42)** чем пять.
  + **(10:44)** Так вот, э-э, если мы потом хотим
  + **(10:50)** просуммировать все эти вот
  + **(10:54)** результаты сравнений, то понятно, что если
  + **(10:57)** никакой связи между… э-э наличием этого признака и принадлежностью объекта к житому классу нет, величина должна быть равная нулю,
  + **(11:05)** получаться.
  + **(11:07)** Э-э, хи-квадрат, мера Пирсона, она этому условию соответствует.
  + **(11:13)** То есть если фактическая частота теоретическая равны, то тогда
  + **(11:16)** разница их равна нулю.
  + **(11:19)** Вот. То есть нет никакой связи между наличием этого итого признака и принадлежностью объекта к житому классу.
  + **(11:25)** Если больше нуля, тогда есть связь, если меньше нуля, тоже есть связь, но отрицательная. То есть этот признак препятствует принадлежности объекта к этому классу.
  + **(11:34)** Вот. А если мы делаем это сравнение путём деления,
  + **(11:38)** тогда
  + **(11:39)** вот это вот N и Т житое делится на теоретическую частоту,
  + **(11:43)** а она равна сумме по строке числа признаков из матрицы абсолютных частот, сумма по колонкам, по колонке, делённая на общее число признаков по всей матрице сопряжённости.
  + **(11:58)** И тогда получается вот это вот выражение.
  + **(12:01)** И для того, чтобы его нормировать к нулю, чтобы потом использовать аддитивный интегральный критерий, то есть потом просто сложить
  + **(12:10)** вот эти
  + **(12:11)** частные критерии и получить
  + **(12:13)** некое одно число, характерное для
  + **(12:16)** класса житого,
  + **(12:22)** можно взять логарифм вот этого отношения, а можно просто вычесть единицу.
  + **(12:27)** Если мы берём логарифм, то это называется мера Харкевича. Если вычитаем единицу, то вы, наверное, в курсе, слышали, это называется коэффициент возврата инвестиций, РОИ. Return on Investment.
  + **(12:41)** Вот. Значит, получается, что в науке, в разных областях, были выработаны уже э-э способы разумного сравнения
  + **(12:52)** этих вот чисел, разные способы.
  + **(12:57)** И если мы посмотрим на вот это отношение,
  + **(13:00)** которое получилось под логарифмом или коэффициент РОИ, то мы увидим,
  + **(13:05)** что это фактически э-э условная относительная частота, делённая на безусловную, потому что N и Т житое,
  + **(13:13)** разделённая на N житое - это равно P и Т житое, это безусловная частота
  + **(13:19)** наблюдения
  + **(13:22)** этого признака у объектов житой категории, житого класса.
  + **(13:26)** А N итое, разделённое на N - это безусловная относительная частота наблюдения этого признака по у объектов всей выборки.
  + **(13:35)** То есть получается, мы что видим?
  + **(13:40)** Что
  + **(13:43)** когда мы сравниваем эти фактическую и теоретическую частоту не путём вычитания, а путём деления, то это у нас получается логарифм условной относительной частоты, делённый на безусловную относительную частоту встречи этого этого признака в житой категории.
  + **(14:01)** Ну и это и есть классическое выражение Харкевича для количества информации,
  + **(14:05)** которую мы получаем, узнав, что у объекта есть итый признак о том, что он относится к житому классу.
  + **(14:13)** Таким образом, вы видите, как интересно получается, что хи-квадрат тесно связано с количеством информации по Харкевичу и тесно связано с коэффициентом возврата инвестиций, который применяется в экономике.
  + **(14:26)** Посмотрим на сами эти модели. Вот эта модель,
  + **(14:31)** посчитанная на основе варианта Харкевича, количество информации.
  + **(14:38)** Мы здесь видим, что вот этот размер под руку, он даёт нам
  + **(14:43)** вот этот признак 0,779 бит информации о том, что это средство связи, и 0,231 бит о том, что это элемент компьютера. Как и должно быть.
  + **(14:54)** Мы знаем, что для средств связи он более характерен, чем для элементов компьютера. И для элементов компьютера тоже количество информации положительное, потому что вероятность встретить этот признак в этом классе объектов тоже выше, чем в среднем.
  + **(15:14)** Вот. И мы видим, что есть числа и отрицательные здесь. Это что означает? Что у каких-то объектов определённой категории есть
  + **(15:24)** некоторые признаки,
  + **(15:26)** которые у них встречаются реже, чем в среднем. И наблюдение такого признака несёт информацию о непринадлежности к данной категории.
  + **(15:34)** То же самое мы видим и в хи-квадрат.
  + **(15:39)** Модель Inf3.
  + **(15:44)** Здесь тоже положительные значения говорят о связи,
  + **(15:48)** наличии связи между наличием этого признака
  + **(15:52)** и принадлежностью классу.
  + **(15:57)** А величина отрицательная тоже о наличии связи, но уже
  + **(16:01)** отрицательной. То есть наличие этого признака несёт информацию или говорит о том, что объект не принадлежит данному классу.
  + **(16:09)** Мы видим, что мера хи-квадрат
  + **(16:11)** у средств связи, вот этот размер под руку, 2,1, а у элементов компьютера - 0,9. То есть практически в два раза больше э-э связь между наличием этого признака и принадлежностью класса
  + **(16:24)** у средств связи, чем у элементов компьютера.
  + **(16:28)** И если эта величина близка к нулю,
  + **(16:32)** то это значит, что практически никакой связи и нет между наличием этого признака и принадлежностью объекта классу. Я вам приводил такой пример когда-то там, наверное, когда рассказывал
  + **(16:43)** э-э про то, как осуществляется распознавание,
  + **(16:46)** что вот если есть у нас там кто-то за дверью, неизвестно кто, и надо установить, кто. И мы можем задавать вопросы ему.
  + **(16:55)** Можем спросить там, допустим: "Какая… волосы длинные?" Нам говорят: "Да". Это признак, который намного чаще встречается у девушек, чем в среднем по выборке. То есть он несёт много информации о том, что это девушки.
  + **(17:08)** А потом мы спрашиваем: "А в брюках?" Нам говорят: "Да, в брюках". А брюки, наличие этого признака несёт маленькое количество информации, небольшое, о том, что это ребята, и небольшое количество информации о том, что это не девушки. Почему небольшое?
  + **(17:23)** Потому что почти все девушки тоже в брюках.
  + **(17:27)** И разница между э-э э-э относительной частотой встречи этого признака в группе ребят и в группе девушек, она мало отличается от средней.
  + **(17:37)** Поэтому получается количество информации очень незначительно в этом признаке. А если бы я ещё спрашиваю: "А мобильный телефон есть?" Мне скажут: "Да, есть".
  + **(17:45)** Это никакой информации не несёт о том, э-э парень это или девушка. Почему? Потому что
  + **(17:51)** этот признак есть у всех э-э с относительной частотой 100% встречается. И во всех группах, и в целом по всей выборке.
  + **(18:02)** То есть получаются три вида признаков: несущих много информации о принадлежности к какому-то классу и непринадлежности к другому. Признаки вот детерминистского типа признаки, но они, может быть, не совсем детерминистские, но близко к этому. А вообще детерминистские детерминистские признаки - это те, которые однозначно говорят о принадлежности объекта классу и встречаются только у объектов этого класса, у объектов другого класса не встречаются.
  + **(18:27)** Ну, есть признаки, которые не совсем так, но почти, почти так, близко к этому. И признаки статистические. Есть близкие к детерминистским и близкие э-э к бесполезным, типа вот признак брюки, вообще бесполезные. И что-то среднее между ценными признаками и не очень ценными.
  + **(18:51)** Вот.
  + **(18:53)** [тишина, фоновый шум]
  + **(18:57)** Какие-то иногда странные эффекты наблюдаются. Какая-то рамочка появляется. Что это значит эта рамочка, не знаете? Я не знаю.
  + **(19:11)** Вот.
  + **(19:13)** Так что модель Inf1, она автоматизирует и другие модели системы когнитивные, автоматизирует сравнение
  + **(19:21)** условных и безусловных относительных частот.

**4. Достоверность моделей**

* **4.1. Необходимость проверки:**
  + **(19:24)** Я как-то раз применил здесь термин "информация", количество информации… Вероятность, вероятность. Вместо термина "относительная частота" применил э-э вероятность. И, соответственно, можно говорить об условной вероятности, безусловной вероятности. Значит, я могу вам сказать, как они связаны. Значит, вероятность - это есть предел, к которому стремится относительная частота при неограниченном увеличении объёма выборки.
  + **(19:53)** Мы когда говорим, то иногда не слишком э-э несём большую ответственность за то, что говорим, даже особо не вдумываемся в то, что говорим, насколько эти слова важны и что они значат.
  + **(20:08)** И вот я когда вам это сейчас вот сказал определение, это определение из теории вероятности просто.
  + **(20:15)** То сразу возникает вопрос: а что понимается под неограниченным увеличением объёма выборки?
  + **(20:22)** Значит, я могу вам сказать, что реально такого неограниченного увеличения объёма выборки быть не может. Всегда она ограничена. Больше того, она довольно жёстко бывает ограничена. То есть часто встречаются небольшие выборки.
  + **(20:35)** Поэтому о вероятности здесь говорить… здесь говорить не стоит. И, во всяком случае, нужно понимать, что когда мы используем термин "вероятность", то мы допускаем некую погрешность в наших высказываниях, отклоняемся от истины, скажем так. Почему?
  + **(20:53)** Потому что относительная частота отличается от вероятности. Между ними есть некоторая разница всегда.
  + **(21:01)** И это вот и есть та погрешность нашего высказывания в случае, если мы применяем термин "вероятность". То есть есть некоторая неточность. Но эта неточность, она стремится к нулю при увеличении объёма выборки.
* **4.2. Затраты на точность:**
  + **(21:15)** Насколько быстро? Ну я вам скажу так, что где-то когда 480 наблюдений, около 400, между… ну, в общем, 400… э-э 400-500 наблюдений, ну около 400. То уже около 5% погрешность между относительной частотой и вероятностью.
  + **(21:32)** Когда 1250 наблюдений - это из статистики известно - при равномерном случайном распределении величины случайной, то такое… такая погрешность наблюдается, так она уменьшается.
  + **(21:45)** При 1250 получается погрешность около двух с половиной процентов.
  + **(21:51)** Различие между относительной частотой и вероятностью.
  + **(21:59)** И потом она стремится к вероятности, но всё медленнее и медленнее. Например, чтобы получить 1% э-э погрешности, нужно уже 10.000 наблюдений.
  + **(22:11)** То есть вы заметили, что сначала она быстро стремится к вероятности, потом всё медленнее и медленнее. И никогда её не достигает. А теперь представьте себе, что если у нас э-э информация стоит денег, то, допустим, мы опрашиваем респондентов и строим какие-то модели на основе этих опросов и анализируем эти модели. То опрос респондента, скажем, стоит рубль, ну, к примеру. Ну я так, на самом деле он рубль не стоит, но пусть, пусть чтоб единичка была. Там 100 рублей или 1000 рублей он стоит.
  + **(22:43)** Вот. Ну давайте, допустим, будем считать, что он рубль стоит. Ну тогда получится так, что для того, чтобы получить информацию с погрешностью 5%, нужно 400 рублей потратить. Чтобы 2%, 2,5 - ну 1000… 1200 рублей, скажем, да? А если мы хотим 1%, то тогда 10.000 нужно потратить, понимаете?
  + **(23:02)** Это уже как бы это самое, начинаешь задумываться, оно надо вот эти 1,5% там ещё получить. 2,5% - 1200 стоит, а 1% стоит э-э 10.000, понимаете?
  + **(23:19)** То есть мы добиваемся того, что получаем ещё 1% э-э точности, ещё на 1% погрешность уменьшается, и платим за это 9.000. И вопрос возникает такой: а стоит этот процент точности 9.000?
  + **(23:34)** Если там мы занимаемся, допустим, экспериментальной физикой, где там у нас 15… девяток после запятой там, допустим, считается достоверным результатом, ну тогда может быть, может быть. Там миллиарды долларов, триллион, может даже и больше вот от большого адронного коллайдера там. А если мы занимаемся социологией, политологией, то этот 1%, может быть, вообще не очень и важен. Хотя вот мы сейчас смотрим на выборы США и видим, что там, в общем-то, сейчас-то там явное различие. А вот выборы были Буш-Гор, там доли процента составили… составило различие в числе голосов. И какой-то ящик, который нашли где-то там под столом, сыграл роль, и Буш стал президентом. Ну так говорят, ну это как шутка такая своего рода своего рода. Ну, в общем, короче говоря, бывает, что это и стоит того, чтобы посчитать это точнее. Но редко так бывает. И обычно люди считают деньги, говорят: "А нет, нам, в принципе, нас устроит и 2,5%, нам больше не надо". Вот. И не платят денег на опрос этих респондентов.
  + **(24:38)** Ну, примерно вот смысл такой во всём этом.
  + **(24:41)** То есть я хочу сказать, что точность стоит денег, понимаете? Чтобы получить более высокую точность, нужно больший объём данных. А их сбор э-э и ввод в систему, обработка - всё это стоит денег. Поэтому возникает вопрос, насколько это необходимо, насколько это оправданы эти затраты.
  + **(25:00)** Вот.
* **4.3. Цель моделей и важность проверки достоверности:**
  + **(25:01)** Теперь, когда мы модели получили, возникает вопрос: а можно ли ими пользоваться этими моделями? Значит, я сейчас вам задам такой вопрос. Ну он такой риторический, сократовский, называется вопрос. То есть я вас спрашиваю, задаю вам этот вопрос не потому, что я не знаю ответа и хотел бы от вас получить ответ. А потому что я хочу вас навести на определённую мысль, к которой вы должны прийти сами, а не просто я вам должен это сказать. Ну если я могу вам это сказать, но лучше, чтобы вы сами к этому пришли. Вот я хочу вас спросить, ребята. Тем более лабораторная работа. Вот, так что вы должны быть в диалоге со мной на этой лабораторной работе. Вы значит, на лекции ещё ладно, значит, там можно слушать там, хотя тоже было бы неплохо, если бы задавали вопросы, хотя бы в конце лекции.
  + **(25:52)** Вот. Ну, вопрос такой: зачем вообще нужна модель? Вот зачем мы создаём эти модели? Вводим данные там, какие-то расчёты проводим. Зачем это нужно?
  + **(26:08)** Кто знает?
  + **(26:13)** М? Кто-нибудь может ответить на этот вопрос?
  + **(26:20)** Ну, наверное, чтобы проверить, как это будет работать в реальности, прежде чем осуществлять?
  + **(26:29)** Ну, это… это ответил ты правильно, только на другой вопрос, не на тот, который я задал сейчас. Это ты ответил на вопрос, э-э зачем проверять достоверность модели, которую сейчас у нас следующая, мы будем рассматривать.
  + **(26:45)** А вообще, зачем она нужна?
  + **(26:49)** Для анализа и прогноза. И чтобы сопоставлять одну ситуацию с другими.
  + **(26:56)** Ну, да, ну это вы конкретно ответили, а можно вот так более обобщённо ответить?
  + **(27:01)** Для решения различных задач.
  + **(27:04)** В частности, действительно, для прогнозирования, сравнения ситуаций, то есть всё это верно. Ну это как просто частные случаи. А эти задачи, которые можно решать, их можно классифицировать. Но я их классифицирую на три группы.
  + **(27:18)** Вот. Эти три группы сейчас я вам на экране покажу.
  + **(27:23)** Вот. Три группы задач.
  + **(27:33)** Ну это моё, по крайней мере, деление, моя классификация. Ну, по-моему, и другие нормальные люди тоже точно так же классифицируют. Аналогично.
  + **(27:43)** То есть есть задачи идентификации, распознавания,
  + **(27:48)** прогнозирования - это один блок задач.
  + **(27:50)** Второй блок - это задачи принятия решений
  + **(27:54)** или управления.
  + **(27:55)** И третья - это решение задач исследования предметной области путём исследования модели.
  + **(28:02)** Вот. Ну для того, чтобы эти задачи решать, каждая из них имеет очень много разных вариантов, в том числе и такие, которые с трудом позволяют их разделить эти задачи.
  + **(28:15)** Вот, допустим, принятие решения тоже можно путём прогнозирования осуществлять, понимаете, путём идентификации или на основе идентификации. То есть это всё переплетается. Потом при принятии решений в простейшем варианте может быть, мы и можем чисто эту задачу решить просто саму по себе. И то она может рассматриваться как задача, обратная задача прогнозирования. Потому что при прогнозировании мы по факторам определяем, что произойдёт с объектом моделирования. А при принятии решения мы наоборот, по целевому состоянию объекта моделирования определяем, какие нужны факторы, чтобы объект перешёл в это состояние. То есть это обратная задача прогнозирования. То есть всё равно она связана с задачей прогнозирования.
  + **(28:56)** Вот. Но в развитой форме принятие решений предполагает уже и использование некоторых результатов исследования предметной области, типа кластерного анализа. Понятно, да?
  + **(29:08)** То есть это мы рассмотрим попозже. А сейчас я могу сказать, что ответ ваш был, в общем-то, в принципе, правильный, что модели нужны для того, чтобы решать задачи.
* **4.4. Достоверность vs. Непроверенность и последствия:**
  + **(29:19)** Ну и теперь вопрос, который вы ответили, хоть я его и не задал, только подумал, а вы уже там… наверное, телепатия у вас или что там развито?
  + **(29:29)** То есть вы забегаете вперёд, э-э по сути дела, м-м предугадываете ход мыслей моих, изложения. Вот. И отвечаете даже на вопросы, которые я ещё не успел задать, а только подумал о них и к ним подвожу, собственно говоря, изложение.
  + **(29:46)** Значит, вот все эти задачи, ребята, чтобы их решать, для этого нужно предварительно убедиться в том, что модель что? Что модель достоверна. То есть, что она адекватно отражает, правильно отражает моделируемую предметную область.
  + **(30:04)** Потому что если это не так,
  + **(30:07)** то мы при идентификации будем ошибаться.
  + **(30:11)** А идентификация - это слово, термин научный, который имеет синонимы: распознавание, классификация, идентификация, диагностика. Представьте себе, в медицине неверно болезнь диагностируют пациенту. И будут лечить его от другой болезни. Ничего хорошего ожидать не приходится при этом, правильно?
  + **(30:34)** Вот.
  + **(30:36)** То есть это чревато такая ошибка э-э нежелательными последствиями. А если мы возьмём недостоверную модель и будем прогнозировать с помощью неё, и потом будем э-э какие-то как-то корректировать свою деятельность с учётом этого прогноза. А модель недостоверная, прогноз такой, который сделан на основе недостоверной модели, скорее всего, не осуществится.
  + **(31:04)** Если осуществится, оно просто случайно так может получиться, что и осуществится. Но скорее всего, что нет.
  + **(31:11)** И тогда получится, что мы будем э-э исходить своих… в своей деятельности из неверного представления о будущем, понимаете?
  + **(31:21)** Вот. Соответственно, наши затраты будут э-э убытками практически, скажем так. Впустую они будут, в лучшем случае, а то ещё и какой-то вред несут.
  + **(31:33)** Ну, по крайней мере, такой вред, что они не принесут пользы.
  + **(31:38)** Потерянная прибыль, как говорят, или потерянная польза.
  + **(31:43)** Вот. А если мы будем принимать решения на основе неверной модели, тогда что у нас будет получаться?
  + **(31:50)** Вот мы хотим э-э оказать воздействие на фирму, чтобы она перешла от состояния э-э предбанкротного в нормальное. Берём эту недостоверную модель, вырабатываем решения соответствующие, рекомендации. И эта фирма нам доверяется и выполняет эти рекомендации. И как вы думаете, у неё получается перейти в нормальное состояние, если модель недостоверна, если она неправильно отражает зависимости между причинами и следствиями в этой предметной области?
  + **(32:22)** Да предприятие вообще может и обанкротиться, я вам скажу. Это весьма рискованно э-э такие давать рекомендации, которые на основе непроверенной модели. Или, скажем так, недостоверной модели.
* **4.5. Разница между непроверенной и недостоверной моделью:**
  + **(32:35)** А теперь будем… хотя это разные вещи: непроверенная модель и недостоверная модель. Сейчас я вам расскажу различия.
  + **(32:44)** Если напомните.
  + **(32:45)** И третья группа задач - исследование моделируемой предметной области.
  + **(32:50)** Вот мы когда исследуем модель э-э и делаем вывод о том, что то, что мы получили выводы какие-то на основе модели, что эти выводы имеют отношение к объекту исследования, объекту моделирования, то это только в том случае, ребята, они имеют отношение к объекту исследования, познания, к объекту познания. Только в том случае, если модель его отражает, если модель правильно отражает объект познания. А если она его не отражает, то у нас нет никаких оснований переносить результаты исследования модели на объект э-э моделирования. То есть мы не можем считать, что если мы такие закономерности или свойства этого объекта выявили на основе модели, что это действительно вот такими свойствами этот объект исследования обладает. Потому что вообще-то модель может его не отражать. Вернее так, она его неверно отражает. Значит, соответственно, и выводы о его свойствах у нас получатся неверные на основе этой модели.
  + **(33:49)** Представляете себе? Мы будем там писать в книжках, что у него такие свойства, а у него нет таких свойств. Потому что модель недостоверна.
* **4.6. Пример с непроверенной моделью:**
  + **(33:59)** А теперь в чём разница между э-э моделью, достоверность которой не изучалась, и недостоверной моделью, ребята?
  + **(34:08)** Значит, модель, достоверность которой не изучалась, не исследовалась,
  + **(34:12)** она может быть как достоверна, так и недостоверна.
  + **(34:16)** И в различной степени.
  + **(34:19)** То есть о ней вообще ничего не известно. В частности, она может быть недостоверной.
  + **(34:23)** Поэтому, если, допустим, идёт какая-то защита, а я присутствовал, как член совета на защитах очень много раз. И могу вам даже привести реальный пример, ну без фамилий там и без это, детализации, но так суть.
  + **(34:39)** Э-э одна девушка защищалась, соискатель учёной степени. И бойко так рассказывала, что она там то-то сделала. Но я скажу так, между нами возникало некое сомнение, что она это сделала, потому что там модель математическая была из книжки приведена э-э из раздела "Массовое обслуживание", теория массового обслуживания. Один к одному просто там формулы были переписаны. А в этой книжке там внизу странички, а мы эту книжку знаем с Валерием Ивановичем, сидим рядышком, который недавно скончался 19 октября. Он эту книжку писал, понимаете? А я по ней преподавал.
  + **(35:18)** Вот. Некоторые лекции за него читал, когда он меня просил. И вот я помню, что внизу этой страницы там есть фраза такая, что обращаем ваше внимание, ну тех, кто читает эту книжку, что эта модель, она не обобщённый характер имеет, то есть она применима не для всех систем,
  + **(35:38)** а только для автоматизированных телефонных станций, для которых она разработана. То есть это модель
  + **(35:44)** массового обслуживания для телефонных станций.
  + **(35:49)** И написано, что на телефонной станции описывает она хорошо. Но, допустим, для супермаркетов она вообще непригодна, или для, допустим, м-м многоядерных процессоров,
  + **(36:01)** многопроцессорных систем, например. Она непригодна для них. Там это прямо написано явно, понимаете, чёрным по белому в книжке в этой.
  + **(36:10)** Вот. А она как раз моделировала э-э супермаркет, систему массового обслуживания. А там внизу написано, что она неприменима для этого, понимаете? Вот эта девушка, значит, эту модель показала, там все это всё покивали, покивали эти профессора. Валерий Иванович слушает, что она несёт? Ты посмотри. Я говорю: "Да вообще, - говорю, - кошмар какой-то". Говорю: "Это ж вообще… Вот на этом месте можно защиту прекращать, понимаете?"
  + **(36:33)** Вот, потому что там ну неверная модель применяется. А дальше там всё на основе модели, все выводы делаются.
  + **(36:40)** Вот. Ну мы сидим все там тихонечко там хихикаем. Так, делаем какое-то умное выражение лица с трудом.
  + **(36:48)** Вот. И тут, значит, вдруг она потом говорит: "Вот мы разработали такую модель", а потом слушайте внимательно. "И выработали вот такие рекомендации. И какая-то фирма там в Адыгее, одна из Адыгеи была, эта фирма применила эти рекомендации". Мы сидим: "Ё-моё!" Мы сидим вообще вот так вот уже э-э за голову хватаемся, думаем: "Так она, во-первых, модель применила, которая некорректна для этих целей применять. А потом, не проверив её достоверность, взяла и выработала рекомендации, эта фирма несчастная, эти рекомендации применила. Представляете себе?"
  + **(37:24)** И что с этой фирмой стало?
  + **(37:29)** Можете себе вообразить?
  + **(37:33)** Ну, единственное, спасает, знаете что? Я вам сейчас скажу, что спасло эту фирму. Знаете? Сказать вам, что её спасло? Давайте, спрашивайте меня, что спасло эту фирму?
  + **(37:43)** А я вам скажу, что спасло.
  + **(37:49)** Ну вы сегодня игры не играете, да? Не общаетесь. Ну её спасло то, что на самом деле никто там ничего не вырабатывал, никто ничего не применял. Никто никому не сообщал никому.
  + **(38:00)** Вот только это её спасло. Вот если бы они реально, действительно, выработали бы рекомендации, но им было это сложно сделать, потому что эту модель надо ещё автоматизировать как-то, чтобы это сделать, чтоб численно посчитать.
  + **(38:12)** Вот если бы они выработали и применили бы эти рекомендации, то этой фирме хана пришла бы, понимаете?
  + **(38:18)** И тут, значит, Валерий Иванович не выдержал и говорит: "Слушайте, вот у меня вопрос есть". Ну это когда уже было обсуждение.
  + **(38:25)** "Вот как вы вообще осмелились применять модель для выработки рекомендаций, э-э когда вы ещё не знаете, достоверна эта модель или нет? Вы представляете себе? Это очень рискованно, это вообще авантюра, понимаете, так между нами говоря. Вот когда революцию совершили октябрьскую в семнадцатом году, кто-нибудь проверял на маленьких странах там, на на кроликах, на мышах, проверял, работает эта модель или нет Маркса-Энгельса-Ленина, или не работает?
  + **(38:58)** Взяли бы, проверили на каком-нибудь там миллиметровой стране, которая на карте не видна. А потом бы уже взяли бы там, масштабировали э-э до страны, которая шестая часть света. Понимаете? Хотя бы так.
  + **(39:13)** Хотя то, что в маленькой стране получается, в большой может не получиться. То, что там э-э в Монте-Карло, например, получается, это во всей России сделать невозможно. Или в Люксембурге. В Люксембурге там высокий уровень жизни. Можно ли сделать в России высокий уровень жизни, пользуясь рецептами, пользуясь рецептами Люксембурга? Нельзя. Почему? Очень большая разница, понимаете? Это Люксембург как юбилейный район примерно, Краснодара. Ну я так уж проутрирую немножко, понимаете?
  + **(39:42)** Так вот,
  + **(39:46)** соответственно, отсюда какой вывод следует? Что, значит, ну Валерий Иванович говорит: "Вы ж вообще-то очень большой риск для этой фирмы создали, потому что ваши рекомендации вообще-то могли бы не быть корректными, неправильными. Потому что, ну про то, что модель некорректна, он не стал говорить, потому что это вообще рушит всю диссертацию. То есть её перечеркнуть можно, выкинуть. Вот. А он говорит: "Хотя бы проверили бы достоверность модели, а потом уже вырабатывали бы решение на её основе, если она достоверна, естественно. А так, говорю, разве можно? Это он говорит, разве так можно делать?" И она говорит: "А мы", - говорит, слушайте внимательно ответ этой соискательницы, который меня тоже шокировал, в общем. Ну я уже к этому науке, уже всё нормально.
  + **(40:36)** Вот. Она говорит: "А вы знаете, вот эта фирма применила наши рекомендации, всё получилось хорошо".
  + **(40:45)** [смешок]
  + **(40:51)** Вы слушаете, что я рассказываю вообще, нет? Ребята? Всё получилось хорошо. И тогда Валерий Иванович уже… ну значит, вам повезло, говорит. Вам просто очень повезло, понимаете? Благодарите Бога, что он вас спас. Хотя вы действовали очень э-э опрометчиво, неразумно, рискованно. Ну вот вам просто повезло, понимаете, раз получилось хорошо. А могло получиться не очень хорошо и вообще нехорошо, понимаете?
  + **(41:18)** Вот и всё. Весь вывод из этой ситуации. То есть это уже называется профонирование, профанация, понимаете? Не наука, а профанация науки. И крайне авантюрные и рискованные действия.
  + **(41:35)** И всё из-за чего? А маленькая такая деталь. Вот этот ромбик,
  + **(41:41)** ромбик, видите? Вот этот ромбик - выбор наиболее достоверной модели.
  + **(41:46)** Вот здесь надо ещё нарисовать
  + **(41:48)** ещё выход вот сюда вот. И нарисовать: "Достоверной модели нет". Вот я дорисовываю, сделаю. Достоверной модели нет вообще. И мы тогда вообще не можем эти задачи решать.

**4.7. Оценка достоверности (введение)**

* **(42:02)** [тишина]
* **(42:08)** Ну, слово "наиболее" убрать. Ну, наверное, даже надо оставить. Вот. Ну, короче говоря,
* **(42:16)** э-э очень важный момент - исследование достоверности модели. Вот к этому вопросу мы сейчас и переходим. То есть если достоверность модели не исследована и неизвестна, то её применение - это вообще профанация и авантюризм, и очень большой риск.
* **(42:34)** А если, значит, она высокая, достаточно высокая, тогда можно ей пользоваться более-менее спокойно, уверенно. Если э-э наша обучающая выборка, на основе которой она создана, репрезентативна по отношению к той генеральной совокупности, где мы собираемся применять модель. То есть вот мы создали модель, а это… на основе исходных данных. А эти исходные данные, они хорошо отражают закономерности в той предметной области, где мы эту модель будем использовать. Где относится, в частности, вот эта фирма. И тогда всё совершенно корректно можно использовать, потому что там действительно те же самые закономерности, которые мы выявили на основе обучающей выборки. А если мы взяли, применили эту модель… Вот взяли, сделали там на данных этих Адыгеи эту модель, а потом взяли на Вашингтонщине там или или Аризонщине где-нибудь там применили.
* **(43:27)** Вот. Или Калифорнийщине.
* **(43:31)** Ну, в общем, вы поняли, да?
* **(43:36)** И вообще там может эта модель ничего не отражает практически того, что там происходит. То есть она не соответствует условиям, допустим, Запада Америки, которые очень сильно отличаются от условий, так сказать, Адыгеи. И получается, что в этом случае э-э применение этой модели там вообще не… неправомерно, является просто чистой авантюрой. А когда наоборот, ребят, наоборот, разработали в США какую-то модель, мы берём её, вот там же вот в MIT разработали или Стэнфорде, дядюшка. А мы берём эту модель и применяем вот здесь вот у нас в Краснодаре. Это что, лучше, что ли?
* **(44:18)** Вот когда они наши модели применяют, мы смеёмся, это полный бред, понимаете, применять наши модели там. Мы это понимаем, осознаём. При этом мы берём их модели и применяем у себя. И не осознаём, что это вообще тоже такой же точно бред. Что это нельзя делать, что нужно свои разрабатывать, отражающие нашу реальность.
* **(44:42)** Значит, я могу сказать, что многие так, конечно, понимают, но помалкивают. А некоторые даже и не помалкивают. Вот и в Перми есть э-э доцент Алексеев Александр Олегович. Он э-э исследовал несколько моделей американские, петербургские модели исследовал, э-э описывающие состояние предприятий строительной отрасли
* **(45:11)** в Перми.
* **(45:13)** В Перми, в Перми.
* **(45:16)** Вот. Оказалось, американская модель вообще э-э на уровне случайного угадывания практически, недостоверно работает.
* **(45:25)** Э-э петербургская модель чуть-чуть получше, но тоже очень плохо.
* **(45:29)** А модель, которая сделана в системе Эйдос, работает великолепно, на уровне 90 там трёх, 92-93% достоверности.
* **(45:39)** Значит, как вы думаете, почему, ребят? Причина, в чём заключается?
* **(45:44)** [тишина]
* **(46:02)** Ребят, когда я спрашиваю, надо отвечать, потому что у нас занятие лабораторное, понимаете? Оно в диалоге должно проводиться. А вы на меня смотрите там…
* **(46:11)** Наверное, потому что взяла американские данные изначально, из-за этого?
* **(46:15)** Американские данные, они совершенно не отражают условия в Перми. Ну, потому что были взяты данные, которые именно соответствуют ситуации в Перми.
* **(46:27)** Конечно, они не являются репрезентативными. Не являются репрезентативными, то есть по отношению к нашей генеральной совокупности. То есть если взять генеральную совокупность, включающую нас, нашу Пермь российскую, то эти данные американские нерепрезентативны. А петербургские? Лучше, нет?
* **(46:48)** Как вы думаете? Петербург и Пермь, они очень похожи, нет? Как вы считаете? Да Петербург больше похож на Нью-Джерси, чем на Пермь. Вы понимаете? Он больше похож на западные города, европейские, чем на российские. Просто действительно объективно он похож на них по своим показателям.
* **(47:14)** Значит, петербургская немножко получше, но не намного, понимаете, не принципиально.
* **(47:19)** И вот я сейчас хочу самое главное сказать. Э-э модель, разработанная в системе Эйдос в Перми, она лучше не потому, что система Эйдос лучше, чем те системы, которые в Петербурге и в Америке используются. Там тоже люди соображают очень неплохо. И тоже они используют хорошие методы, корректные и сильные. И тоже у них есть программные системы. То есть свет клином не сошёлся на системе Эйдос в этом смысле. То есть есть и системы получше, чем система Эйдос. Я знаю это, уверен в этом абсолютно. Ну, может быть, не во всех отношениях, но лучше, но в некоторых. У системы Эйдос есть одно очень важное достоинство. Знаете какое? Может оно смешно немножко это звучать. Ну догадайтесь, какое у неё достоинство. Почему она имеет некоторое преимущество перед другими системами?
* **(48:22)** Никогда не догадаетесь. Я так вижу, что вы не можете догадаться. Ну разве что вот Ирина может догадаться. Или София, или Марина. Или Дима. Или, может, Виктория догадается. Ну Артур точно догадается. Я просто Артур просто он просто обязан догадаться.
* **(48:47)** У неё нет ничего лишнего?
* **(48:49)** Нет, она просто бесплатная. Ну да.
* **(48:57)** Значит, просто можно взять её и воспользоваться, понимаете? Применить её для решения своих задач. А эти вот американские системы самые дешёвые системы, аналогичного функционального назначения, 150.000 долларов стоят.
* **(49:10)** Вот возьми её и воспользуйся в Перми. Понимаете? Это нереально. Даже для Пермского национального университета. Это просто, ну, идиотизм.
* **(49:22)** Откуда я эти цифры взял? Я показывал систему Эйдос в управлении президента, э-э управлении делами президента Российской Федерации, замначальнику управления делами, профессору Райкову.
* **(49:36)** И он говорит: "Так вы что, хотите сказать, что вы сделали систему э-э лучше американских?" Я говорю: "Да нет, не хочу". Вот. "А за сколько вы хотите её продать?" Я говорю: "Да я вообще-то не по этому пришёл". Вот. Ну вот он тогда и говорит: "Самые дешёвые американские системы, аналогичные функциональной системе Эйдос, стоят 150.000 долларов. Самые дешёвые, простенькие. А получше системы, они стоят там и 300, и 400, и 800. Вот Wall Street, допустим, система, она 800.000 долларов стоит. Просто система программная. Понимаете?
  + **(50:16)** Вот и всё.
  + **(50:17)** То есть, э-э доцент Алексеев подумал… думал он очень недолго, я вам скажу. Или заплатить 150.000 долларов за самую паршивую американскую систему интеллектуальную, решающую такие задачи. Или скачать с моего сайта систему Эйдос и попробовать. Он думал, знаете сколько? Одну сотую секунды. Может, даже одну тысячную секунды он думал. Он вообще не думал, он взял, скачал и попробовал. И получилось достоверность гораздо выше, чем у этих систем там Wall Street, там и всё прочее, понимаете? Почему? Потому что если бы взяли систему Wall Street и применили в Перми, то, конечно, бы получилось лучше, чем в системе Эйдос, я уверен. Но её там не применили, понимаете?
  + **(51:08)** Вот такая вот маленькая деталь. Её там не применили и не применят никогда.
  + **(51:16)** Вот и всё.
  + **(51:17)** Вот такое у неё достоинство есть. Ну а второе достоинство, оно не настолько серьёзное, но тоже важное. Она работает реально. То есть она вполне разумно сделана, убедительно, аргументированно, модели, там всё это вполне обосновано.
  + **(51:35)** И самое главное, что это всё работает, действительно, понимаете? Работоспособно.
  + **(51:40)** В тех областях, где есть эксперты, они говорят: "Да, так это ж мы и так знаем". Вот. А в тех областях, где их нет, мы становимся экспертами, проведя исследования и узнав, какие там зависимости, мы оказываемся теми, кто лучше всего разбирается в этой предметной области. И разобрались мы в этой предметной области новой, в которой ещё нет экспертов, благодаря системе Эйдос. Она выявляет зависимости, очень хорошо их э-э в различных формах представляет, в том числе и аналитической форме. Я вам это не показывал, но она и для регрессионного анализа готовит данные. Вот, то есть всё это там есть.
  + **(52:19)** Почему? А возникала такая необходимость в этом, и я это сделал. Просто это не уже не уровень лабораторных работ, это уровень диссертации уже.
  + **(52:29)** Вот и всё.
* **4.8. Как измеряется достоверность:**
  + **(52:31)** Поэтому мы переходим к вопросу о том, как измеряется достоверность. И хочу ваше внимание вот на что обратить. Как только где-нибудь когда-нибудь в будущем, в будущем, ребята, вы услышите слово "модель", вот кто угодно, вот из вас, где-нибудь на какой-нибудь защите там, дипломной, диссертации, э-э кто-нибудь услышит… здесь даже доклад какой-нибудь на конференции, услышит слово "модель". Кто-то там что-то вы там всё сидели, там что-то с телефончиком, там что-то нажимали. Вдруг услышали, там "модель" сказали. Выступающий сказал слово "модель". Вы сразу должны оторваться и подумать, вспомнить то, что я сейчас вам буду говорить. Э-э интересно, вот у вас должна появиться такая мысль: интересно, он сейчас скажет про достоверность модели или нет? Этот докладчик? Или в статье вы прочитаете. Если нет, ребята, если нет, вот вы можете спокойно абсолютно закрывать на этом месте эту статью, доклад, ничего там дальше интересного не будет. Почему? Потому что человек писал недалёкий, понимаете? Не понимающий элементарных вещей. Причём человек, может быть, даже опасный, потому что если он что-нибудь там рекомендует кому-нибудь на основе своей модели, достоверность которой он не проверил, то неизвестно, чем это кончится, понимаете? Может, его потом с собаками будут искать этого рекомендатора. Я серьёзно говорю совершенно. То есть люди такие скажут: "А вот он нам дал такую рекомендацию, и мы потерпели вот такой-то убыток. А ну давайте-ка найдём его и накажем его за это", - понимаете?
  + **(54:08)** Вот в чём дело.
  + **(54:10)** Это же не смешно, когда там предприятие может потерпеть убытки, ну многомиллионные, скажем так. А можно было дом построить, а можно было вот рекомендацию эту применить.
  + **(54:22)** Вот выбирайте или рекомендацию, или дом. Применили рекомендацию, всё, дом вы не построите. Потом нашли этого человека, который эту рекомендацию вам дал, и сказали, что вы про него думаете, понимаете? Так прямо конкретно.
  + **(54:36)** Это же серьёзно всё, это же не игрушки.
  + **(54:44)** Поэтому вы должны, как только услышите слово "модель", сразу вспоминать про слово "достоверность", сразу же. И должны поинтересоваться: а проверяли её достоверность или нет? Если нет, то о чём вообще можно говорить тогда? О каких… никаких решений, о каких задачах? Да никаких. Вообще нельзя пользоваться этой системой, этой моделью. Это крайне опасно, и это авантюра и профанация науки, если кто-то построил модель и, не проверив её достоверность, начал применять. Как у нас в семнадцатом году.
  + **(55:16)** Применили. Ну, получилось что-то такое, что сейчас спорят все с переменным успехом. Надо было, конечно, на кроликах попробовать или на миллиметровых странах проверить сначала, которые не видно на карте. Вот было бы тогда более корректно.
  + **(55:34)** Хотя эти миллиметровые страны, они меня сейчас услышали, о них сказали: "А почему на нас? Вы тогда на сантиметровых странах проверяйте, это более будет похоже вот на вашу страну. То есть как-то лучше будет отражать условия вашей страны". Я говорю: "Ну давай". Тогда давай на километровых странах, которые полглобуса занимают, с собой сторон видно страну глобуса. Ну так вот и сделали. Ну и так оно и получилось, что очень много уехало, очень много погибло. Потом ещё очень много сами поубивали. Потом война. Ну, в общем, очень тяжёлая история. И я думаю, что, конечно, это был очень рискованный эксперимент такой социальный, крупномасштабный.
  + **(56:22)** Ну нужно было иметь очень большую смелость, чтобы пойти на такой эксперимент. Без проверенной модели, ребята. Она была хорошо обоснована, логично, убедительно Марксом. Ну и Энгельс, и Ленин, вроде бы всё очень хорошо обосновали, но потом оказалось, что там не совсем всё понятно. Ленин э-э вообще высказался таким образом, что мы построили не совсем то, а точнее совсем не то, что собирались. То есть он признал, что получилось не то совсем, публично, понимаете? Потом предложил НЭП. НЭП - это было, конечно, гениальная совершенно идея. Вот то, что сейчас в Китае - это вот и есть НЭП, то, что Ленин предложил. У нас он длился недолго, до смерти Ленина. А потом пришёл Иосиф Виссарионович и всё стало как раньше.
  + **(57:15)** Вот. Ну и что мы получили? А если бы НЭП продолжался, то у нас было бы ещё круче, чем в Китае. Ничего бы не распалось, Советский Союз бы. И сейчас бы мы были бы первой страной мира, наверное, я так думаю. Потому что в Китае это же самое происходило, но через 50 лет там, через 60 после НЭПа. То есть у нас был бы огромный задел во времени. И надо не надо забывать, что мы строили в Китае заводы там, и технологии им давали все. И они ещё и сами их брали иногда. И так они со всего мира не против взять. Ну, в общем, была бы другая история абсолютно, понимаете?
* **4.9. Мера достоверности Ван Рейсбергена (введение):**
  + **(57:58)** Вот. Значит, теперь давайте посмотрим, как измеряется достоверность модели. Значит, для этого мы переходим в режим 3-4.
  + **(58:06)** Вот. И здесь вот у нас есть Help
  + **(58:09)** в этом режиме.
  + **(58:11)** Который сейчас я вам этот Help в чат пошлю.
  + **(58:21)** Ну это не сам Help, а режим
  + **(58:25)** оценки достоверности модели. Вот и Help этого режима.
  + **(58:31)** Да, ребята, те, кто пользуется системой Эйдос, должны знать, что, значит, я изменил доступ к сайтам системы, э-э к сайту обновлений, в ТП-серверу. И с целях безопасности. Вот. И поэтому обновить систему нужно, используя режим 6.2, а в нём есть пункт шесть. Просто скачиваете оттуда обновление, патч, разворачиваете в папочке системы, когда система не запущена, и у вас система обновляется до текущей версии, которая 24 июня. То есть позавчерашняя.
  + **(59:14)** Понятно, да?
  + **(59:17)** Вот. И ещё сейчас одну вам информацию дам,
  + **(59:21)** интересную, раз уж зашло речь.
  + **(59:29)** [тишина]
  + **(1:00:07)** Быстро пополняется в облаке.
  + **(1:00:12)** Вот. Значит, э-э теперь смотрим сам этот Help.
  + **(1:00:19)** Ну, и я уже задавал вопрос: что должна делать модель? И вы ответили правильно. Вот. Но вопрос возникает такой: а как оценить достоверность модели? А вот как раз вот так и оценить: решаем задачи и смотрим, получается или нет. Если получается, значит модель достоверна. Если не получается, значит недостоверна.
  + **(1:00:38)** Но только эти задачи надо решать на ретроспективных данных. То есть не прямо вот сейчас вживую вырабатывать рекомендации и кому-то давать их сразу же, а и потом смотреть, что с ним произойдёт, и будет ли он вас потом ловить там и искать, понимаете? А нужно взять ретроспективные данные. Ну, допустим, фондовый рынок Forex. Берём за уже прошедший период э-э этот рынок, временные ряды. И попытаемся на основе прошлых показателей этого рынка э-э прогнозировать будущие показатели. Получается - значит модель достоверна, отражает закономерности этого рынка.
  + **(1:01:18)** Ясно, да? Принцип.
  + **(1:01:20)** То есть тоже э-э проверяется достоверность путём решения задач. Тогда возникает вопрос: а какие задачи лучше всего брать для этой цели?
  + **(1:01:30)** Я вам сказал, что есть задачи трёх типов: задача идентификации, задача принятия решений и задача исследования.
  + **(1:01:39)** Значит, я вам скажу так, что самой простой из этих всех задач, задача, значит, идентификация, прогнозирование, принятие решений и исследование, вот так точнее можно сказать. Задача прогнозирования очень похожа на задачу идентификации, только у них э-э при идентификации свойства объектов там, признаки их и принадлежность классам каким-то относятся к одному периоду времени. Вот признаки сейчас и принадлежность классу тоже сейчас. А при прогнозировании факторы действуют в прошлом,
  + **(1:02:09)** а результат их действия в будущем относительно
  + **(1:02:13)** времени действия факторов.
  + **(1:02:16)** [тишина]
  + **(1:02:20)** То есть это уже более сложная задача поэтому. Ну, если есть ретроспективные данные, то, в общем-то, математика вся и алгоритмы одинаковые решения этих задач, просто разная интерпретация, к какому времени относятся э-э признаки и к какому времени относятся классы.
  + **(1:02:38)** И вот мы можем это же сделать и путём исследования модели, то есть принятия решений. Тоже принимаем решение, смотрим, что произойдёт. И совпадает, не совпадает. Вот. Ну вот как при… как э-э проверять достоверность модели на задачах принятия решений на ретроспективных данных, не совсем понятно. Там же уже решение принято, и уже последствия э-э реализовались этих решений.
  + **(1:03:08)** А вот что касается идентификации и прогнозирования, там более-менее понятно. То же самое непонятно, как можно э-э измерять достоверность модели путём решения задачи исследования объекта моделирования. Ну, взяли там его, исследовали, какие-то сделали выводы на основе модели, я имею в виду. Потом стали сравнивать с фактической стороной дела, оно не совпадает. Ну значит, модель недостоверная.
  + **(1:03:35)** Ну это всё такие какие-то экзотические способы проверки э-э с использованием этих сложных задач. А самым простым является проверка э-э достоверности модели путём решения задачи идентификации. Значит, есть такой учёный Ван Рейсберген, который это понял, то, что я сейчас говорю, очень давно. Ну, если честно, я не знаю, когда настолько давно, ну лет 50 назад, наверное, он это понял так примерно.
  + **(1:04:05)** Вот.
  + **(1:04:07)** Можно попробовать узнать.
  + **(1:04:16)** [тишина]
  + **(1:04:31)** Боюсь, что не найдём мы, когда она предложена.
  + **(1:04:59)** [тишина]
  + **(1:05:22)** Вот. Короче говоря, э-э довольно давно это предложено, эта мера. Вот этот Ван Рейсберген, я не знаю, сейчас он жив или нет, но он в очень преклонном возрасте, я в Википедии смотрел информацию о нём.
  + **(1:05:36)** Значит, э-э
  + **(1:05:41)** что он предложил, собственно говоря? Он э-э задался таким вопросом: что должна делать модель при решении задачи идентификации? И ответил на него очень просто: она должна обеспечивать
  + **(1:05:56)** определение принадлежности объектов идентификации к каким-то классам, к которым она относится,
  + **(1:06:03)** и непринадлежность тем классам, к которым этот объект не относится. То есть определять принадлежность объекта к тем классам, к которым он относится, и непринадлежность тем, к которым он не относится.
  + **(1:06:15)** И понятно, что модель может и ошибаться.
  + **(1:06:19)** То есть э-э система, работающая на основе модели, может и ошибаться. То есть она может неверно определять принадлежность объекта классу. Ну то есть она определяет, что он принадлежит классу, а он на самом деле не принадлежит. И неверно определять непринадлежность классу. То есть она определяет, что он не принадлежит этот объект некоторому классу, а на самом деле он к нему принадлежит.
  + **(1:06:42)** Ну то есть она должна правильно определять принадлежность классу, правильно определять непринадлежность классу. И может ошибаться во всех этих случаях, и в первом, и во втором случае.
  + **(1:06:52)** Это называется ошибками первого-второго рода.
  + **(1:06:57)** Как же посчитать э-э, в общем, реально достоверность модели? Если у нас есть некая выборка,
  + **(1:07:05)** распознаваемая,
  + **(1:07:07)** состоящая из определённого числа объектов, описанных признаками,
  + **(1:07:13)** и мы эти объекты идентифицируем, относим к тем или иным классам или и не относим,
  + **(1:07:19)** то у нас какое-то количество решений будет истинным, какое-то решение… количество решений будет ложным.
  + **(1:07:26)** Решения бывают положительные и отрицательные. Кроме того, что они истинные и ложные, ещё есть положительные и отрицательные.
  + **(1:07:33)** Положительные решения - это решения, когда уровень сходства больше нуля объекта с классом. Поэтому они и называются положительными. И
  + **(1:07:44)** эти положительные решения, позитив, они могут быть и истинными (True), и ложными (False). То есть можно вот сказать так, что есть э-э True Positive и False Positive решения.
  + **(1:07:59)** И есть также решения отрицательные - это решения э-э когда уровень сходства объекта с классом меньше нуля. Это решение о непринадлежности объекта классу. И они тоже есть истинные и ложные. Есть True Negative решение - это истинное решение о непринадлежности. То есть система предположила, что не принадлежит, и действительно не принадлежит к этому классу объект. И есть False Positive… False Negative. Что это значит? Что система определила, что объект не принадлежит классу, а он на самом деле принадлежит, понимаете?
  + **(1:08:36)** Соответственно, Ван Рейсберген, я так предполагаю, что скорее всего, он, наверное, программист, э-э ну, по крайней мере, не чужд программирования, потому что он предложил ввести четыре сумматора. Э-э название которых он предложил такие: ТП, ТН, ФП, ФН, то есть True Positive, True Negative, False Positive, False Negative. Значит, почему я думаю, что он программист? Потому что
  + **(1:09:00)** э-э математики, они обозначают переменные одной буквой и ещё там могут быть индексы. Ну, индексы - это не название переменной, а это какие-то э-э, скажем так, тоже переменные, означающие какие-то аргументы. Вот, аргументы, то есть это принадлежность этой переменной каким-то элементам каких-то списков, массивов, вот так примерно. Вот. Так вот, м-м, раз он обозначил эти переменные двумя буквами, то остаётся вариант такой, что он либо… то есть она не характерна для математиков вот такие обозначения. Э-э значит, он либо программист, либо экономист. Экономисты тоже этим балуются, то есть они прямо берут иногда целыми словами обозначают переменные. В принципе, это как в программировании делается. В программировании именно так и делается.
  + **(1:09:54)** Вот. Ну, скорее всё-таки, что он программист, чем экономист, потому что э-э всё-таки его работа математическая. То есть там ряд формул был выведен. Из них я самые простейшие вам покажу, а остальные несложно найти в интернете. Я и статьи написал про эти тематики, там ссылки для источников. Ну их полным-полно, можно найти просто вот поиском.
  + **(1:10:22)** Значит, когда идентифицируется один объект распознаваемой выборки или тестовой выборки, как в нейронных сетях говорят. Но мне этот термин "тестовая выборка" не нравится, потому что он предполагает, что мы тестируем модель. А мы можем не тестировать, а просто использовать модель, понимаете, для решения задач. Конечно, использование модели тоже является в каком-то смысле её тестированием. То есть если она всё время она успешно работает на практике, то это тоже подтверждает, что она хорошая. Вот. Но мы-то цель ставили не тестировать модели, а использовать. А тестируем только в случае, когда мы хотим достоверность её определить. Тогда да, тогда мы тестируем. Ну вот, мы-то меру достоверности используем не только тогда, когда мы тестируем её, но и когда просто пользуемся, потому что тоже её используем для этих… Ну ладно.
  + **(1:11:09)** Короче говоря, что предложил Ресберген? Он предложил ввести вот эти сумматоры. И когда любой объект тестовой выборки этой вот распознаваемой выборки, лучше говорить, более общее понятие, э-э идентифицируется с классами,
  + **(1:11:25)** с обобщающими категориями, то какой-то из этих сумматоров обязательно увеличивается на единичку. Потому что решение, любое решение, которое примет система о принадлежности, непринадлежности, оно может быть либо истинным, либо ложным, либо положительным, либо отрицательным. Других вариантов никто пока не придумал ещё. И в том числе и я тоже, пока других вариантов не придумал.
  + **(1:11:48)** Ну, возникает э-э м-м вопрос такой: что потом с этим делать, с этими сумматорами? Ресберген предложил формулы для точности модели и полноты модели, которые в Help'е здесь вот они приводятся эти формулы. А потом с использованием этих понятий точность и полнота модели, он, по-моему, среднее геометрическое взял. Вот. И это э-э обозначил как F-меру. Почему F, я не знаю, почему он F назвал её. Вот, ну может, полная мера, там Full какой-нибудь там. Вот, или Фаня там.
  + **(1:12:27)** Вот. Ну, в общем, короче говоря, F-мера достоверности модели Ван Рейсбергена.
  + **(1:12:36)** Вот. Ну теперь давайте посмотрим, насколько эта мера хороша, насколько она подходит для системы Эйдос.
  + **(1:12:44)** Система Эйдос - это мультиклассовая система, в которой каждый объект выборки,
  + **(1:12:50)** и обучающий, и распознаваемый, можно относиться ко многим классам одновременно.
  + **(1:12:58)** Ну, я вам приведу пример. Вот да, а в мере Ван Рейсбергена предполагается, что объект относится к одному классу, и к другим никаким не относится. Ну я скажу так, что это он, конечно,
  + **(1:13:14)** ну, слишком уж сильное такое предположение. Это очень сильное предположение, которое фактически не соответствует действительности. Значит, если взять, допустим, меня, к примеру, то я э-э кто? Преподаватель вуза, профессор, дедушка, отец, муж, сын, брат,
  + **(1:13:44)** понятно, да? И так далее, и так далее. Житель города Краснодара, россиянин, землянин, ну то есть я с Земли. Хотя недавно здесь, но всё-таки вот сейчас я бы сказал так, что всё-таки я с Земли.
  + **(1:14:01)** Вот. Ну это шутка такая. Ну как недавно? Уже так довольно давно, да? Вот сколько уже возраст какой. Вот. Так вот,
  + **(1:14:10)** то есть я, ну, совершенно очевидно, что я отношусь ко многим категориям одновременно, обобщающим, они более общие, чем я, чем конкретная личность там, профессор Луценко, да?
  + **(1:14:21)** Вот. То есть эти категории имеют более общий характер, и я к ним отношусь.
  + **(1:14:27)** А почему-то Ван Рейсберген думал, что этого не может быть. Ну как-то странно он думал. А как тогда быть, вот если объект относится ко многим классам, а мы хотим эту меру применить? Я придумал такой вариант, что можно считать, что объект, э-э любой объект состоит из э-э тождественных объектов, их назвал логические объекты. Этот физический объект, я вот физический объект, да, вот конкретно меня видно там и так далее. А логически я состою из там десятков, сотен точно таких же одинаковых объектов, которые ничем от меня не отличаются от физического, но один относится к классу э-э мужчины, другой к классу муж, третий к классу брат, четвёртый к классу дедушка.
  + **(1:15:11)** То есть столько логических объектов, сколько классов.
  + **(1:15:17)** И я являюсь суперпозицией этих логических объектов, но поскольку они тождественны, то эта суперпозиция эквивалентна исходному объекту.
  + **(1:15:26)** Вот. Но логически можно рассматривать, что я отношусь ко многим классам вот именно таким образом. То есть если, допустим, у нас какая-то есть база данных, где указано, к какому классу относится Евгений Вениаминович Луценко, то там можно сделать строчек 20 в этой базе данных, в этой таблице и указать там в одной строчке, что я к одному классу отношусь, в другом, к другому. А объект называется Евгений Луценко. То есть это будет означать, что это один и тот же объект. То есть эти все строчки, они будут сгруппированы э-э по этому вот, по моему имени отчеству. А информация о том, что я ко многим классам отношусь, она вот таким образом может быть представлена.
  + **(1:16:08)** Даже это более правильно с точки зрения нормализации баз данных, вот подходов к базам данных, даже более правильно именно вот так и делать, чем писать в одной строке, к каким классам я отношусь, какие у меня признаки. Хотя, надо признать, в одной строке удобнее. Вот. Ну есть некоторые ограничения, правда, то есть там, допустим, на число этих перечислений, может быть, ограничения и так далее. А когда вот мы строки добавляем, то там такие ограничения есть на размеры баз данных, но они гораздо слабее.
  + **(1:16:40)** [тишина]
  + **(1:16:46)** Смотрим на следующие моменты. То есть, короче, с этой вот мультиклассовостью нужно разбираться. Следующее, смотрим на результаты идентификации.
  + **(1:16:58)** Система Эйдос представляет собой систему, которая
  + **(1:17:02)** определяет не просто принадлежность объекта классу, а ещё определяет степень уверенности своей в этом решении.
  + **(1:17:11)** Вот, допустим, она, видите, пишет, что мышка один, объект конкретный, относится к классу мышка, и она уверена в этом на 90 там 2,68%. И ещё она считает, что этот объект относится к классу элемент компьютера, в этом она уверена на 72%. И фактически так и есть. Вот эти птички стоят там, где это фактически действительно э-э это так и есть. Ещё она думает, что это может быть относится к классу телефон. Но в этом она уверена только на 48%. А ещё она думает, что это может быть даже и монитор, но только на 28, на 29% она в этом уверена. И она уверена в том, что это не вешалка, не сумка, не аксессуар, не мебель, не спорт-инвентарь.
  + **(1:17:55)** И вот так вот мы можем посмотреть на разные объекты. И вот мне нравится клавиатура 2. Смотрите, клавиатура 2, она относится к классу клавиатура на 95%, к классу элемент компьютера на 56%. А к классу аксессуары, сумка относится на 4%. Это неверно, она к ним на самом деле фактически не относится, то есть птички здесь нет. Это ложно-положительное решение.
  + **(1:18:19)** Значит, это ошибка системы, ложная идентификация. Но ошибка небольшая. Почему? Потому что она сама система думает, что всего лишь на 4% этот объект относится к этому классу. Значит, что значит относится на 4%, на 20%, на 90%? Что это вообще значит? А это значит, что этот объект похож именно настолько похож на вот этот класс.
  + **(1:18:48)** Значит, я думаю, что можно считать, что распознавание - это разложение в ряд. А вопрос возникает: разложение в ряд чего и ряд по чему, по каким функциям? Я предложил считать, что функция, описывающая объект, вектор объекта, разлагается в ряд по векторам классов. А в качестве весовых коэффициентов разложения в ряд используются вот эти вот степени уверенности.
  + **(1:19:23)** Значит, это всё подробно описано вот в этих статьях. В Research Gate описано с численным примером, а в журнале только теория описана. А численный пример будет в отдельной статье будет, которая сейчас вот в размещаемом номере будет.
  + **(1:19:38)** Это очень глубокая, очень интересная, э-э такое интересное э-э как сказать, представление о том, что такое распознавание.
  + **(1:19:48)** И вот мы здесь видим, смотрите, ребята, что, видите, э-э система определила, в соответствии с интегральным критерием, который я потом вам расскажу попозже, когда будем про задачи рассматривать. Сейчас мы рассматриваем достоверность, о том потом будем подробно рассматривать задачи. Там я расскажу про интегральные критерии, про их виды там и так далее. Так вот, здесь мы видим что, что система, она ошиблась, отнесла сумку к аксессуарам. Но она ошиблась немного, совсем чуть-чуть, всего 4% уверенность самой системы в том, что это так. То есть она относит, но очень сильно сомневается в этом. Всё-таки она относит к этому, к этой категории, а не к категории стул. Стул вообще не относят, видите, с минусом знак, да? То есть уровень сходства. А здесь она относит, но с очень низким уровнем сходства. Что делает Ван Рейсберген? Он бабах, и к сумматору True Positive… False Positive, то сложно-положительное решение, суммирует, чтобы вы думали, он единицу суммирует к сумматору. Он подсчитывает сами решения, ложные, истинные, положительные, отрицательные. Сколько таких решений? Я говорю: "Ну ты, Ваня, э-э Ресберген, прямо ну крутой ты какой-то совершенно, прямо человек, очень строгий, понимаете? Вот, и несправедливый, я считаю. Зачем ты так наказываешь систему и модель, суммируя единицу к сумматору False Positive, когда обоснованно, оправдано суммировать 4 сотых, ребята, не единицу, а 4 сотых. Потому что 4% - это это 4 сотых. Она же сама, говорю, не уверена в своём решении. Она уверена только на 4 сотых. Ну давайте и суммировать к сумматору False Positive 4 сотых вот эту саму уверенность. И тогда мы увидим, что система, да, она ошибается, но она ошибается чаще на э-э решениях с очень низким уровнем сходства. И мы это увидим сейчас, ребята.
  + **(1:21:48)** Вот смотрите. Берём мы
  + **(1:21:53)** строим график частотных… частотное распределение
  + **(1:22:04)** решений истинных, числа… числа истинных и ложных решений положительных и отрицательных. Истинные решения красным обозначены, ложные решения синим. Справа от оси Y - положительные решения, слева от оси Y - отрицательные решения.
  + **(1:22:24)** Замечательный чат в системе Teams.
  + **(1:22:31)** И смотрим, решение отрицательное, о непринадлежности объекта классу. Они… а здесь ложных и нет, здесь все истинные. Все решения истинные. А решение о принадлежности есть и истинные, и ложные. Ну здесь сами эти решения, а здесь сглаженная кривулька, чтобы было видно смысл этих вот э-э самой исходной кривой. Когда данных побольше э-э тестовой выборки, тогда здесь таких изломов нет, она тогда очень похожа на нормальное распределение, два нормальных распределения. А здесь очень маленькая выборка, всего 20 примеров, поэтому вот так выглядит. Значит, теперь смотрите, ребята, значит, ложные решения есть? Да, есть, но они есть при маленьких уровнях сходства.
  + **(1:23:41)**