ПРОГРАММА ИНДИВИДУАЛЬНОГО КУРСА «МАШИННОЕ ОБУЧЕНИЕ В R, PYTHON И H2O 6.2»

Форма занятий – индивидуально онлайн (скайп или Zoom). Индивидуально, не в группе!
Стоимость занятий – 2000 руб./1 час (график – 2 занятия по 2 часа), 2400 руб./1 час (график – 1 занятие по 1 часу, 2 занятия по 1 часу в неделю).
Продолжительность – 36 часов.
График занятий – 2 занятия по 2 часа, 2 занятия по 1 часу в неделю, 1 занятие по 1 часу.
Цены действуют с 15 декабря 2019 года. Все договоренности до этой даты – в силе!
Слоты:

- понедельник – с 17:00 по 23:00
- вторник – с 10:00 по 12:00
- среда – с 21:00 по 23:00
- четверг – с 17:00 по 23:00
- пятница – с 10:00 по 12:00, с 19:00 по 23:00
- суббота – с 13:00 до 15:00, с 21:00 по 23:00*

* обратите внимание, некоторые слоты могут быть заняты учениками предыдущих итераций курса.

Объем материала по языкам: Python – 85%, H2O – 10%, R – 5%.

Модуль 1. Предварительная подготовка данных

I. Вводная часть

I.1. Типы данных
I.2. Типы переменных
  I.2.1. Количественная переменная
  I.2.2. Категориальная переменная
I.3. Функция, производная, частная производная, градиент, градиентный спуск

II. Знакомство с Python

II.1. Установка Anaconda
II.2. IPython и Jupyter Notebook
II.3. NumPy
II.4. SciPy
II.5. matplotlib
II.6. pandas
II.7. scikit-learn
  II.7.1. Понятие массива признаков и массива меток

VII. Загрузка данных
VIII. Удаление бесполезных переменных, переменных «из будущего», нестабильных переменных
IX. Преобразование типов переменных
X. Нормализация строковых значений
XI. Обработка дублирующихся наблюдений
XII. Обработка редких категорий
XIII. Появление новых категорий в новых данных
XIV. Импутация пропусков
XIV.1. Способы импутации количественных и бинарных переменных
XIV.2. Способы импутации категориальных переменных
XV. Обработка выбросов
XVI. Описательные статистики
XVI.1. Среднее, медиана и мода
XVI.2. Квантиль
XVI.3. Дисперсия и стандартное отклонение
XVI.4. Корреляция и ковариация
XVI.5. Получение сводки описательных статистик в R
XVI.6. Получение сводки описательных статистик в библиотеке pandas
XVII. Нормальное распределение
XVII.1. Знакомство с нормальным распределением
XVII.2. Построение гистограммы и графика квантиль-квантиль для подбора преобразований, максимизирующих нормальность
XVII.3. Подбор преобразований, максимизирующих нормальность
XVII.3.1. Обратное преобразование, отрицательное обратное преобразование
XVII.3.2. Логарифм
XVII.3.3. Логарифм с нормированием на среднее, log(x/mean(x)+k), где k – значение, близкое к 0 или 1
XVII.3.4. Корень четвертой степени
XVII.3.5. Кубический корень
XVII.3.6. Квадратный корень
XVII.3.7. Экспоненциальное преобразование
XVII.3.8. Квадратный корень разности между константой и исходным значением переменной
XVII.3.9. Логарифм разности между константой и исходным значением переменной
XVII.3.10. Возведение в степень
XVII.3.11. Арксинус
XVII.3.12. Преобразования Бокса-Кокса и Йео-Джонсона
XVII.3.13. Пример подбора оптимального преобразования из ранее рассмотренных для переменной с правосторонней асимметрией

XVIII. Проверка статистических гипотез

XIX. Конструирование признаков

XIX.1. Статическое конструирование признаков, исходя из предметной области

XIX.1.1. Поиск сильных переменных

XIX.1.2. Агрегаты

XIX.1.3. Создание переменной, у которой значения основаны на значениях исходной переменной

XIX.1.4. Создание бинарной переменной на основе значений количественной переменной

XIX.1.5. Создание переменной, у которой каждое значение - среднее значение количественной переменной, взятое по уровню категориальной переменной

XIX.1.6. Объединение нескольких бинарных переменных в одну количественную переменную

XIX.1.7. Вычисление расстояния между двумя точками по географическим координатам (через формулу гаверсинусов)

XIX.1.8. Геохеширование

XIX.1.9. Восстановление координат по адресу

XIX.1.10. Выделение из дат единиц времени

XIX.1.11. Учет цикличности временных признаков

XIX.1.12. Макроэкономические переменные

XIX.1.13. Агрегаты (на примере банковских транзакций)

XIX.1.14. MCC-коды

XIX.1.15. Индикатор платежной дисциплины

XIX.2. Статическое конструирование признаков, исходя из особенностей алгоритма

XIX.2.1. Дамми-кодирование (One-Hot Encoding)

XIX.2.1.1. Дамми-кодирование по методу неполного ранга

XIX.2.1.2. Дамми-кодирование по методу полного ранга

XIX.2.1.3. Дамми-кодирование с помощью функции get_dummies() библиотеки pandas

XIX.2.1.4. Дамми-кодирование с помощью класса OneHotEncoder

XIX.2.2. Кодирование контрастами (Effect Encoding)

XIX.2.3. Присвоение категориям в лексикографическом порядке целочисленных значений, начиная с 0 (Label Encoding)

XIX.2.4. Кодирование частотами (Frequency Encoding)

XIX.2.5. Кодирование вероятностями (Likelihood Encoding)

XIX.2.5.1. Кодирование простым средним значением зависимой переменной

XIX.2.5.2. Кодирование простым средним значением зависимой переменной по схеме leave-one-out

XIX.2.5.3. Кодирование простым средним значением зависимой переменной по схеме K-fold

XIX.2.5.4. Кодирование средним значением зависимой переменной, сглаженным через сигмоидальную функцию

XIX.2.5.5. Кодирование средним значением зависимой переменной, сглаженным через сигмоидальную функцию, по схеме K-fold
XIX.2.5.6. Кодирование средним значением зависимой переменной, сглаженным через параметр регуляризации
XIX.2.5.7. Кодирование средним значением зависимой переменной, вычисленным по «прошлому» (упрощенный вариант кодировки, применяющейся в библиотеке CatBoost)
XIX.2.6. Присвоение категориям в зависимости от порядка их появления целочисленных значений, начиная с 1 (Ordinal Encoding)
XIX.2.7. Присвоение категориям, отсортированным по процентной доле наблюдений положительного класса зависимой переменной, целочисленных значений, начиная с 0 (еще одна схема Ordinal Encoding)
XIX.2.8. Бинарное кодирование (Binary Encoding)
XIX.2.9. Бинарное кодирование с хешированием (Hashing)
XIX.2.10. Создание переменных-взаимодействий (interactions)
XIX.2.11. Биннинг переменных
XIX.2.11.1. Биннинг на основе интервалов, созданных вручную или одинаковой ширины
XIX.2.11.2. Биннинг на основе децилей
XIX.2.11.3. Биннинг на основе WoE и IV
XIX.2.11.4. Биннинг на основе CHAID

XIX.3. Динамическое конструирование признаков, исходя из особенностей алгоритма

XIX.3.1. Преобразование категориальных признаков в количественные внутри библиотеки CatBoost
XIX.3.2. Биннинг категориальных признаков внутри библиотеки H2O
XIX.3.3. Связывание взаимно исключающих признаков внутри библиотеки LightGBM

XX. Стандартизация
XXI. Логический контроль
Best practice по предподготовке. Задача StateFarm

Модуль 2. Метрики качества и валидация

I. Метрики для оценки качества модели

I.1. Бинарная классификация
I.1.1. Отрицательный и положительный классы, порог отсечения
I.1.2. Матрица ошибок
I.1.3. Правильность (accuracy)
I.1.4. Чувствительность (sensitivity)
I.1.5. Специфичность (specificity)
I.1.6. 1 – специфичность (1 – specificity)
I.1.7. Точность (Precision)
I.1.8. Сбалансированная правильность
I.1.9. Сравнение точности и чувствительности (полноты)

I.1.10. F-мера
I.1.11. Варьирование порога отсечения
I.1.12. Коэффициент Мэттьюса (Matthews correlation coefficient или MCC)
I.1.13. Каппа Коэна (Cohen's cappa)
I.1.14. ROC-кривая (ROC curve) и площадь под ROC-кривой (AUC-ROC)
I.1.15. Недостатки AUC-ROC
I.1.16. PR-кривая (PR curve) и площадь под PR-кривой (AUC-PR)
I.1.17. Применение PR-кривых и ROC-кривых в условиях дисбаланса классов
I.1.18. Коэффициент Джини (Gini coefficient)
I.1.19. Логистическая функция потерь (logloss)

I.2. Регрессия
I.2.1. R^2, коэффициент детерминации (R-square, coefficient of determination)
I.2.2. Скорректированный R^2, скорректированный коэффициент детерминации (adjusted R-square, adjusted coefficient of determination)
I.2.3. Среднеквадратичная ошибка (mean squared error, MSE)
I.2.4. Корень из среднеквадратичной ошибки (root mean squared error, RMSE)
I.2.5. Средняя абсолютная ошибка (mean absolute error, MAE)
I.2.6. Сравнение RMSE и MAE
I.2.7. Корень из среднеквадратичной логарифмической ошибки (root mean squared logarithmic error, RMSLE)
I.2.8. Средняя абсолютная ошибка в процентах (mean absolute percentage error, MAPE)
I.2.9. Симметричная средняя абсолютная ошибка в процентах (symmetric mean absolute percentage error, SMAPE)

II. Недообучение, переобучение и обобщающая способность
III. Стратегии валидации и доверительный интервал
III.1. Случайное разбиение на обучающую и тестовую выборки
III.2. Обычная *k*-блочная перекрестная проверка
III.3. Повторная *k*-блочная перекрестная проверка
III.4. *k*-кратное случайное разбиение на обучающую и тестовую выборки (перекрестная проверка Монте-Карло)
III.5. Перекрестная проверка с исключением по одному
III.6. Перекрестная проверка, учитывающая группы связанных наблюдений
III.7. Комбинированная проверка для настройки гиперпараметров
III.8. Вложенная перекрестная проверка
III.9. Разбиение на обучающую и тестовую выборки, учитывающее временную структуру данных
III.10. Перекрестная проверка, учитывающая временную структуру данных (перекрестная проверка расширяющимся/скользящим окном)

III.11. Комбинированная проверка для настройки гиперпараметров, учитывающая временную структуру данных
III.12. Необходимость гэпа между обучающей и тестовой выборками в рамках стратегий проверки для временных рядов, реализации стратегий проверки с гэпом
III.13. Перекрестная проверка, учитывающая группы связанных наблюдений, в сочетании с перекрестной проверкой скользящим окном для данных, сочетающий временной ряд и формат «один клиент – несколько наблюдений»
III.14. Бутстреп
III.15. Знакомство с доверительным интервалом
III.15.1. Асимптотический метод
III.15.2. Бутстреп-метод
III.16. Доверительный интервал метрики качества (на примере AUC)
III.16.1. Асимптотический метод
III.16.2. Бутстреп-метод

IV. Сравнение моделей машинного обучения с помощью статистических тестов

IV.1. Сравнение двух классификаторов с помощью критерия МакНемара
IV.2. Сравнение обобщающей способности нескольких классификаторов с помощью *Q*-критерия Кохрена

Модуль 3. Логистическая регрессия

I. Математический аппарат логистической регрессии

I.1. Понятие модели бинарного выбора
I.2. Вероятность, шанс и отношение шансов
I.3. Логит или натуральный логарифм шансов
I.4. Интересные свойства логита
I.5. Переход от логита к шансам, от шансов к вероятности

II. Оценивание коэффициентов логистической регрессии (кейс 1, задача прогнозирования возникновения глубокой просрочки Give Me Some Credit)

II.1. Регрессионный коэффициент
II.2. Экспоненциальный коэффициент

III. Интерпретация коэффициентов логистической регрессии

IV. Предпосылки для построения логистической регрессии

IV.1. Линейная связь между предиктором и логарифмом шансов
IV.2. Единый масштаб измерения переменных
IV.3. Отсутствие мультиколлинеарности
IV.3.1. Диагностика мультиколлинеарности
IV.3.2. Способы борьбы с мультиколлинеарностью
IV.3.3. Регуляризация
IV.3.3.1. Лассо
IV.3.3.2. Гребневая регрессия
IV.3.3.3. Эластичная сеть

V. Важности признаков в логистической регрессии
VI. Специальные метрики для оценки качества модели логистической регрессии
VI.1. Аппроксимации R2
VI.2. Критерий Хосмера-Лемешова
VI.3. AIC и BIC

VII. Знакомство с классом LogisticRegression библиотеки scikit-learn и классом H2OGeneralizedLinearEstimator библиотеки h2o
VII.1. Параметры класса LogisticRegression
VII.1.1. Общие параметры обучения модели
VII.1.2. Гиперпараметры регуляризации
VII.1.3. Настройка оптимизатора
VII.2. Параметры класса H2OgeneralizedLinearEstimator
VII.2.1. Общие параметры обучения модели
VII.2.2. Гиперпараметры регуляризации
VII.2.3. Настройка оптимизатора

VIII. Построение модели логистической регрессии с помощью класса LogisticRegression библиотеки scikit-learn на данных, не содержащих/содержащих временной ряд по схемам line-by-line и pipeline
VIII.1. Поиск оптимальных значений гиперпараметров для последовательности моделей на отложенной выборке
VIII.2. Построение последовательности моделей с оптимальными значениями гиперпараметров на всей исторической выборке
VIII.3. Применение последовательности моделей с оптимальными значениями гиперпараметров, обученной на всей исторической выборке, к новым данным
IX. Построение модели логистической регрессии с помощью класса H2OGeneralizedLinearEstimator библиотеки h2o на данных, не содержащих/содержащих временной ряд по схемам line-by-line и pipeline
IX.1. Поиск оптимальных значений гиперпараметров для последовательности моделей на отложенной выборке
IX.2. Построение последовательности моделей с оптимальными значениями гиперпараметров на всей исторической выборке
IX.3. Применение последовательности моделей с оптимальными значениями гиперпараметров, обученной на всей исторической выборке, к новым данным

X. Работа с дисбалансом классов
X.1. Две основные стратегии: присвоение весов и семплинг
X.2. Удаление примеров мажоритарного класса (undersampling)
X.2.1. Случайное удаление примеров мажоритарного класса (Random Undersampling)

X.2.2. Удаление примеров мажоритарного класса по определенным правилам (связи Томека)

X.3. Увеличение числа примеров миноритарного класса (oversampling)

X.3.1. Случайное дублирование примеров миноритарного класса (random oversampling)

X.3.2. Случайное дублирование примеров миноритарного класса (random oversampling)

X.3.3. SMOTE (Syntetic Minority Oversampling Technique – оверсемплинг за счет создания синтетических примеров миноритарного класса)

X.4. Модель единичных весов

X.5. Делфт-модели

Модуль 4. Деревья решений

I. Знакомство с методом деревьев решений

I.1. Описание метода

I.2. Визуализация работы деревьев решений

I.3. Краткое знакомство с CHAID и CART

I.4. Задачи деревьев решений

I.5. Области применения деревьев решений

I.6. Преимущества деревьев решений

I.7. Недостатки деревьев решений

II. Методы CHAID и XCHAID

II.1. Подробное описание алгоритма CHAID

II.2. Иллюстрация работы CHAID на конкретном примере

II.3. Построение модели дерева CHAID в пакете R CHAID

III. Метод CART

III.1. Подробное описание алгоритма CART

III.2. Принцип неоднородности

III.3. Метод отсечения ветвей на основе меры стоимости-сложности с перекрестной проверкой

III.4. Иллюстрация работы CART на конкретном примере

III.5. Важности предикторов

III.6. Построение модели дерева классификации CART в пакете R rpart

III.7. Построение модели дерева регрессии CART в пакете R rpart

III.8. Построение модели дерева классификации CART с помощью класса DecisionTreeClassifier питоновской библиотеки scikit-learn

III.9. Построение модели дерева классификации CART с помощью класса DecisionTreeRegressor питоновской библиотеки scikit-learn

Модуль 5. Случайный лес

I. Знакомство с методом случайного леса

I.1. Общая идея

II.2. Построение ансамбля деревьев регрессии
II.3. Поиск оптимальных значений гиперпараметров случайного леса с помощью пакета caret
II.4. Улучшение качества логистической регрессии с помощью признаков на основе правил, сгенерированных случайным лесом (пакет R inTrees)

III. Улучшение интерпретируемости случайного леса с помощью пакета R RandomForestExplainer

III.1. Оценка важности предиктора с точки зрения минимальной глубины использования
III.2. Альтернативные метрики важности
III.3. Многомерные графики для оценки важности предикторов
III.4. Парные графики для оценки корреляций между метриками важности
III.5. Графики взаимодействий между переменными
III.6. Получение отчета по построенному случайному лесу
III.7. Тепловая карта прогнозов
III.8. Улучшение качества логистической регрессии с помощью признаков на основе правил, сгенерированных случайным лесом (пакет R RandomForestExplainer)

IV. Построение модели случайного леса в пакете R ranger – быстрой реализации случайного леса

IV.1. Построение случайного деревьев классификации
IV.2. Построение случайного леса деревьев вероятностей
IV.3. Построение случайного леса деревьев выживаемости

V. Построение модели случайного леса в питоновской библиотеке scikit-learn

V.1. Построение случайного леса с помощью классов RandomForestClassifier и RandomForestRegressor библиотеки scikit-learn
V.2. Построение полностью рандомизированных деревьев с помощью классов ExtraTreesClassifier и ExtraTreesRegressor библиотеки scikit-learn
V.3. Построение модели распределенного случайного леса с помощью класса H2ORandomForestEstimator пакета h2o
V.4. Поиск оптимальных значений гиперпараметров случайного леса с помощью классов GridSearchCV и H2OGridSearch
V.5. Построение изолирующего леса с помощью класса IsolationForest библиотеки scikit-learn
V.6. Пакет treeintrepreter для выполнения декомпозиции прогнозов случайного леса (на основе анализа путей решений)
V.7. Улучшение интерпретируемости случайного леса с помощью пакета lime
V.7.1. Три принципа: интерпретируемость, локальная верность и независимость от структуры объясняемой модели

V.7.2. Математический аппарат
V.7.3. Применение для задачи регрессии
V.7.4. Применение для задачи классификации

Модуль 6. Градиентный бустинг

I. Знакомство с методом градиентного бустинга

I.1. Общая идея бустинга
I.2. Связь градиентного спуска с алгоритмом бустинга
I.3. Градиентный бустинг
I.4. Особенности градиентного бустинга
I.5. Настройка гиперпараметров
I.5.1. Общая схема настройки гиперпараметров
I.5.2. Поиск компромисса между количеством итераций и темпом обучения (низкий темп при постепенном увеличении числа итераций, большое число итераций при постепенном снижении темпа)
I.5.3. Глубина
I.5.4. Минимальное количество наблюдений для разбиения узла
I.5.5. Минимальное количество наблюдений в терминальном узле
I.5.6. Максимальное количество терминальных узлов
I.5.7. Биннинг количественных предикторов
I.5.8. Биннинг категориальных предикторов
I.5.9. Способ обработки категориальных предикторов
I.5.10. Параметры формирования подвыборок наблюдений и подпространств признаков: количество наблюдений для построения дерева, доля случайно отбираемых столбцов при построении дерева, доля случайно отбираемых столбцов при формировании каждого уровня дерева, доля случайно отбираемых столбцов при формировании каждого узла дерева
I.5.11. Регуляризация
I.5.12. Ранняя остановка
I.6. Выбор функции потерь
I.7. Выбор типа бустинга
I.8. Схемы настройки гиперпараметров «кольцо» (поиск оптимального соотношения между темпом обучения и количеством итераций, настройка дизайна деревьев с помощью глубины, минимального количества наблюдений в терминальном узле, внесение дополнительной рандомизации за счет настройки подвыборок наблюдений и подпространств признаков, новый поиск оптимального соотношения между темпом обучения и числом итераций)
I.9. Ансамблирование бустингов
I.10. Преимущества и недостатки градиентного бустинга

II. XGBoost

II.1. Общее знакомство
II.2. Регуляризированная целевая функция

II.3. Оценка качества структуры дерева
II.4. Алгоритмы поиска оптимальной точки расщепления
II.4.1. Точный жадный алгоритм поиска оптимальной точки расщепления
II.4.2. Аппроксимационный алгоритм поиска оптимальной точки расщепления
II.4.3. Алгоритм поиска оптимальной точки расщепления на основе гистограммирования
II.5. Поиск оптимальной точки расщепления, адаптированный для работы с разреженными данными
II.6. Архитектура
II.6.1. Блочная структура для параллельного обучения
II.6.2. Кэшируемый доступ
II.6.3. Блоки для вычислений во внешней памяти
II.6.4. Сжатие блоков
II.6.5. Шардирование блоков
II.7. Параметры и гиперпараметры
II.7.1. Общие параметры обучения
II.7.2. Основные гиперпараметры для регулировки сложности
II.7.3. Гиперпараметры, отвечающие за структуру деревьев
II.7.4. Гиперпараметры для рандомизации
II.7.5. Гиперпараметры, настраивающие регуляризатор
II.7.6. Прочие гиперпараметры
II.8. Особенности подготовки данных перед использованием XGBoost
II.9. Обучение модели XGBoost и вычисление важностей признаков (частоты использования, охвата и выигрыша) с помощью нативного интерфейса (функции train() библиотеки xgboost) и интерфейса scikit-learn (классов XGBClassifier и XGBRegressor)
II.10. Вычисление вероятностей классов (задача бинарной классификации)
II.11. Вычисление вероятностей классов (задача регрессии)
II.12. Пакет BoostARoota: отбор признаков с помощью XGBoost

III. CatBoost
III.1. Вводная часть
III.1.1. Предварительное получение расщеплений
III.1.2. Преобразование категориальных признаков в количественные
III.1.3. Выбор структуры дерева
III.1.3.1. Симметричные (небрежные) деревья
III.1.3.2. Тип бутстрепа (присвоение весов наблюдениям)
III.1.4. Вычисление значений листьев
III.2. Обучение модели CatBoost с помощью нативного интерфейса (класса Pool библиотеки catboost) и интерфейса scikit-learn (классов CatBoostClassifier и CatBoostRegressor)
III.3. Перекрестная проверка
III.4. Выбор наилучшей модели

III.5. Ранняя остановка
III.6. Контрольные точки (снапшоты)
III.7. Важности признаков
III.8. Получение прогнозов и сохранение модели
III.9. Настройка гиперпараметров с помощью класса GridSearchCV библиотеки scikit-learn, библиотеки Hyperopt и метода .grid_search() библиотеки catboost
III.10. Параметры и гиперпараметры
III.10.1. Общие параметры обучения
III.10.2. Основные гиперпараметры для регулировки сложности
III.10.3. Гиперпараметры, отвечающие за структуру деревьев
III.10.4. Гиперпараметры для рандомизации
III.10.5. Гиперпараметры, настраивающие регуляризатор
III.10.6. Гиперпараметры, ответственные за предварительное вычисление расщеплений
III.10.7. Гиперпараметры, настраивающие преобразование категориальных признаков в количественные
III.10.8. Гиперпараметры, настраивающие детектор переобучения

IV. LightGBM
IV.1. Вводная часть
IV.1.1. Односторонний отбор на основе градиентов (Gradient-based One-Side Sampling или GOSS)
IV.1.2. Связывание взаимно исключающих признаков (Exclusive Feature Bundling или EFB)
IV.1.3. Повершинное построение деревьев
IV.1.4. Обработка категориальных признаков (one-hot-кодирование, обработка «как есть», превращение в вещественные, поиск оптимального расщепления среди категорий на основе подхода Уолтера Фишера)
IV.2. Параметры и гиперпараметры
IV.2.1. Общие параметры обучения
IV.2.2. Основные гиперпараметры для регулировки сложности
IV.2.3. Гиперпараметры, отвечающие за структуру деревьев
IV.2.4. Гиперпараметры для рандомизации
IV.2.5. Гиперпараметры, настраивающие регуляризатор
IV.2.6. Прочие гиперпараметры
IV.3. Особенности подготовки данных перед использованием LightGBM
IV.4. Обучение модели LightGBM и вычисление важностей признаков с помощью нативного интерфейса (функции train() библиотеки lightgbm) и интерфейса scikit-learn (классов LGBMClassifier и LGBMRegressor)
IV.5. Вычисление вероятностей классов (задача бинарной классификации)
IV.6. Вычисление вероятностей классов (задача регрессии)

О преподавателе:

Занятия ведет Артем Владимирович Груздев, директор ИЦ «Гевисста», переводчик бестселлеров – книги Райан Митчелл «Скрапинг веб-сайтов с помощью Python» https://www.ozon.ru/context/detail/id/136423991/ и книги Андреаса Мюллера и Сары Гвидо «Введение в машинное обучение с помощью Python» https://www.ozon.ru/context/detail/id/140891479/, автор книги «Прогнозное моделирование в IBM SPSS Statistics, R и Python. Деревья решений и случайный лес» https://www.ozon.ru/context/detail/id/142702694/, автор книги «Изучаем pandas» https://www.ozon.ru/context/detail/id/149717036/, автор более трех десятков статей по прогнозному моделированию.

Исследовательский центр «Гевисста» с 2009 г. осуществляет разработку, валидацию, внедрение и мониторинг риск-моделей, моделей оттока, моделей отклика на базе IBM SPSS Statistics, IBM SPSS Modeler, SAS Enterprise Miner, SAS Enterprise Guide, R, Python. Осуществляет подготовку специалистов в сфере прогнозного моделирования и анализа данных. Основное направление – разработка новых высокоточных и одновременно интерпретируемых алгоритмов машинного обучения. Клиентами являются Citibank N.A., TransUnion, DBS Bank, Banco Galicia, StateFarm.