НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ МЕТОДЫ И ИНСТРУМЕНТЫ ДЛЯ АНАЛИЗА ЭКОНОМИЧЕСКОЙ ИНФОРМАЦИИ

Кузин Сергей Сергеевич
директор по консалтингу «Тринити Солюшнс», к.т.н.

Анализ данных

- Статистический анализ и Data mining
- Большие данные
- Процессы Data mining

Процесс Data mining

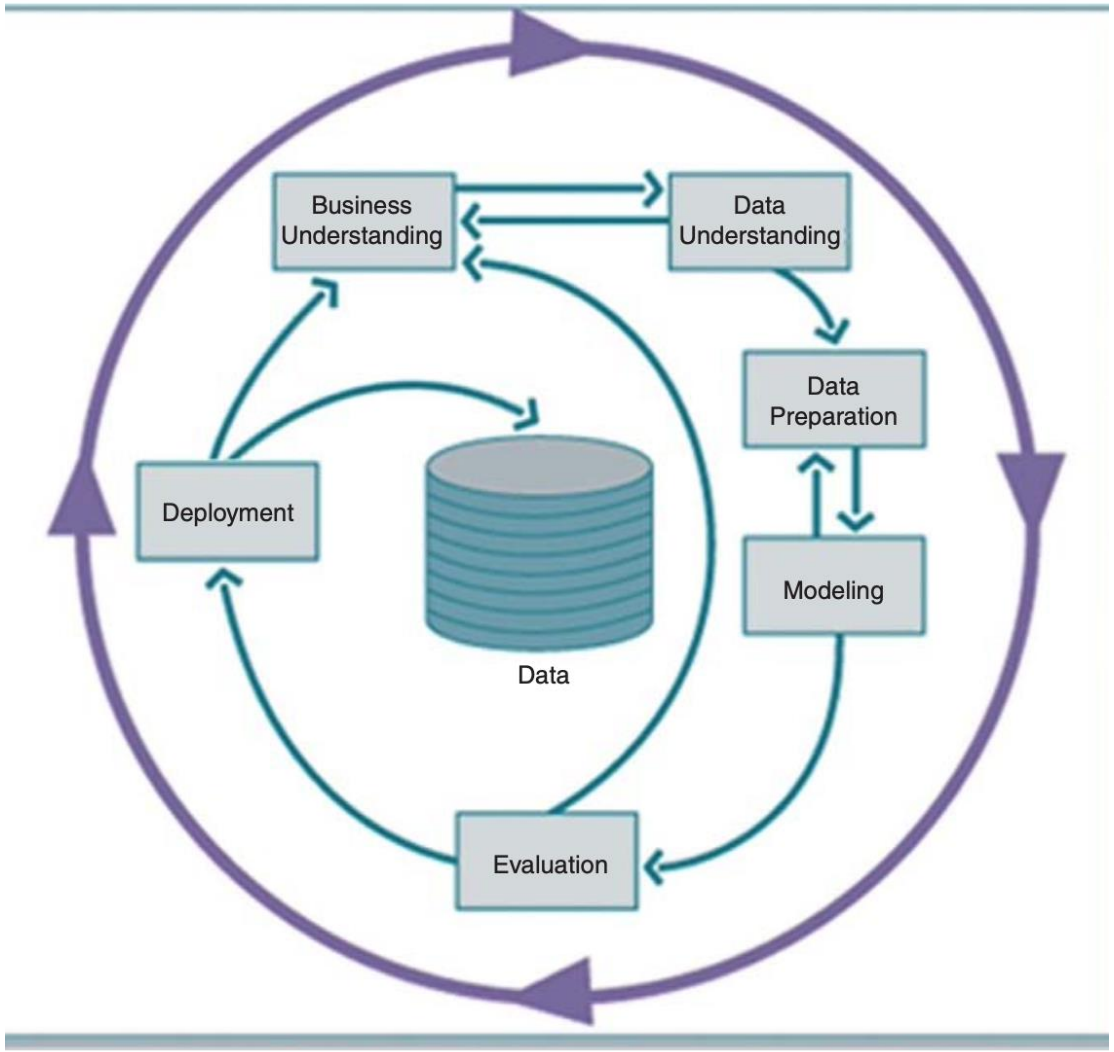
Data mining – это не технический процесс, а процесс решения бизнес-задач. Решение Data mining начинается с формулирования целей, за которым следует сбор необходимых и доступных данных и поиск скрытых закономерностей в данных, помогающих достижению поставленных целей.

Существующее программное обеспечение Data mining может использоваться бизнес-аналитиками, а это означает, что задачи выполняются специалистами, обладающими содержательными знаниями в предметной области. Это обеспечивает высокую степень использования бизнес-знаний в процессе решения задач, что существенно повышает качество решений.

Существуют различные варианты описания процесса Data mining, например, CRISP-DM (Cross Industry Standard of Processes of Data Mining), SEMMA (Sample, Explore, Modify, Model and Assess), DMAIC (Define, Measure, Analyze, Improve and Control). В наиболее полной форме процесс описан в Межотраслевом стандарте процессов Data Mining CRISP-DM.

Этапы процесса Data mining

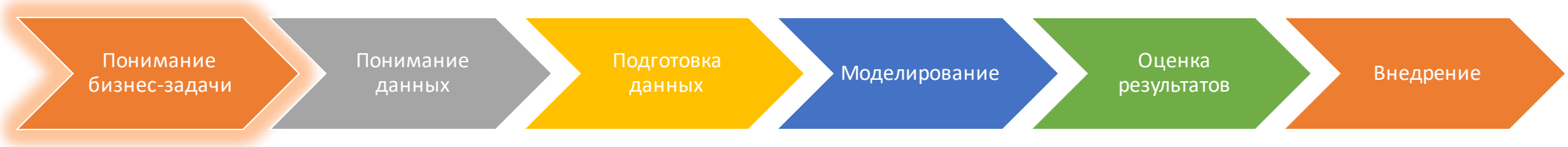
1. Понимание бизнес-задачи
2. Понимание данных
3. Подготовка данных
4. Моделирование
5. Оценка полученных результатов
6. Внедрение



Понимание бизнес-задачи

Понимание целей проекта и требований с точки зрения перспективы использования результатов.

Постановка задачи и подготовка предварительного плана достижения целей.



Понимание данных

Первоначальное определение состава данных, которые могут потенциально доступны и могут быть полезны в решении задачи.

Понимание данных, выявление возможных проблем с качеством данных.

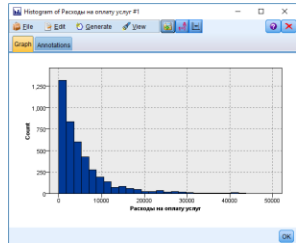
Первоначальный разведочный анализ данных. Выявление подмножеств данных, которые потенциально могут содержать скрытую информацию для решения задачи проекта.

Получение данных

- Доступ к данным
- Интеграция данных
- Первичный отчет по данным

Описание данных

- Переменные, наблюдения
- Описательные статистики



Оценка качества данных

- Пропущенные значения
- Выбросы
- Отчет по качеству данных

Подготовка данных

Этап подготовки данных включает все операции по подготовке финального набора данных из исходных первичных данных для анализа выбранными инструментами. Включает операции отбора данных, контроля и чистки, реструктурирования, агрегирования до необходимого уровня представления, объединение дынных из различных источников и таблиц.

Основными операциями подготовки данных являются подключение к данным, преобразование и отбор данных с целью создания набора данных в формате подходящем для работы с ними аналитическими инструментами.

Вопросы этапа подготовки данных

Чистка данных

Преобразование данных, формирование нужных переменных

Что делать с пропущенными значениями? Импутация данных

Все ли наблюдения вносят одинаковый вклад в анализ? Взвешивание и балансировка данных

Что делать с выбросами? Фильтрация данных

Как работать с временными рядами? Представление данных

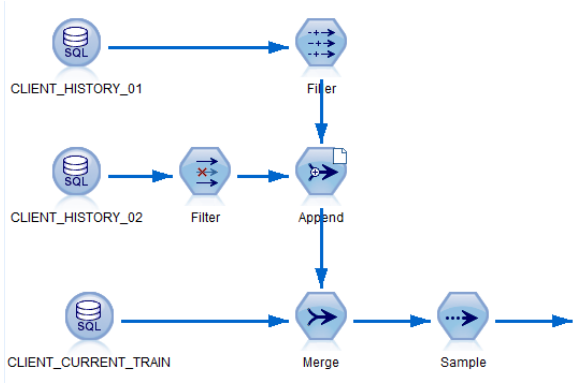
Создание новых переменных

Можно ли сократить объем данных? Записи – выборки данных; переменные – сокращение размерности данных; значения – категоризация, перекодировка

Получение и интеграция данных

Важным этапом подготовки данных является их интеграция. Данные часто представлены в различных форматах или на различных уровнях агрегирования. Существенной частью работы по интеграции данных является мэппинг данных.

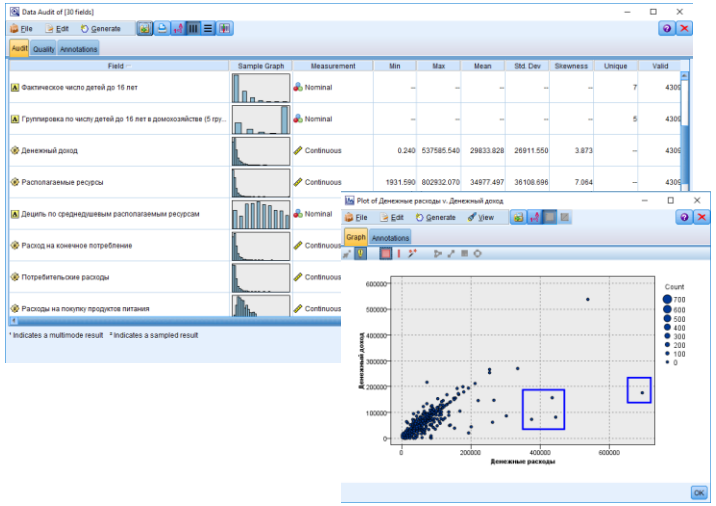
Мэппинг – это описание преобразования исходных данных из первичных источников в данные, непосредственно используемые в решении задачи.



Описание данных

Описание данных – это не просто подготовка списка доступных данных, это аудит данных, просмотр выборки записей по каждой переменной, чтобы иметь представление, что это за данные.

В программных средствах анализа данных имеются различные средства получения описательных статистик по переменным или сводного отчета по аудиту данных с описательными статистиками, отчету по пропущенным значениям и выбросам.



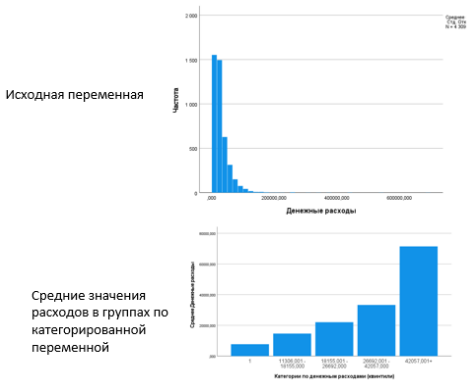
Понимание бизнес-задачи → Понимание данных → Подготовка данных → Моделирование → Оценка результатов → Внедрение

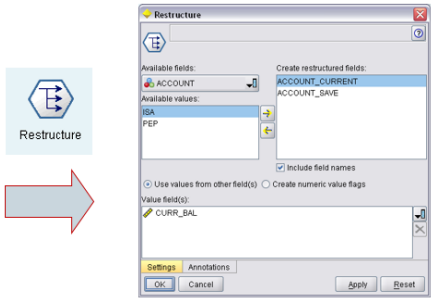
Преобразование данных

Агрегирование данных

Реструктуризация

Преобразование данных, создание новых переменных





Понимание бизнес-задачи | Понимание данных | Подготовка данных | Моделирование | Оценка результатов | Внедрение

Выбор методов моделирования

- Выбор алгоритмов моделирования. Процесс подготовки данных в какой-то мере зависит от выбранного метода/алгоритма моделирования.
- Как правило, для решения поставленной задачи, подходят несколько методов моделирования, поэтому на данном этапе можно не ограничиваться выбором одного конкретного метода, а использовать несколько методов и делать выбор по результатам сравнения точности и устойчивости моделей. Возможно также одновременное применение нескольких моделей и получение результата комбинированием результатов нескольких моделей.
- Предположения о данных. Многие методы моделирования основываются на некоторых предположениях относительно данных, например, нормальность распределения переменных или отсутствие пропущенных значений.

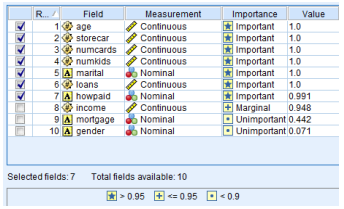
Подготовка выборки данных для настройки и тестирования модели

- Использование случайных выборок для настройки модели в случае большого объема данных
- Разделение выборки на обучающую и контрольную. В некоторых случаях выделение также валидационной выборки, в случае если по результатам проверки модели на контрольной выборке вносятся изменения в некоторые параметры модели.
- Балансировка выборки. В задачах классификации выборка часто балансируется. Например, если в исходных данных категории отклика да/нет представлены неравномерно, скажем, 98% «нет», то для повышения качества настройки модели рекомендуется сбалансировать обучающую выборку, обеспечив примерно равное соотношение категорий.

Отбор предикторов в модели

- Предварительный отбор предикторов. Первоначально в состав предикторов включается широкий перечень потенциально полезных переменных, после чего производится предварительный отбор предикторов в модель путем оценки степени связи каждого потенциального предиктора с целевой переменной (Feature selection). Большое количество предикторов может приводить к неустойчивости модели, особенно при относительно малом объеме выборки.
- При ручном отборе предикторов при настройки моделей ориентируются на оценки статистической значимости предикторов в модели.
- Многие алгоритмы моделирования, содержащиеся в современных программных средствах анализа данных, включают встроенные средства отбора предикторов в ходе настройки модели. В некоторых методах моделирования, например, в деревьях решений отбор предикторов является естественным свойством самого метода.

Feature Selection



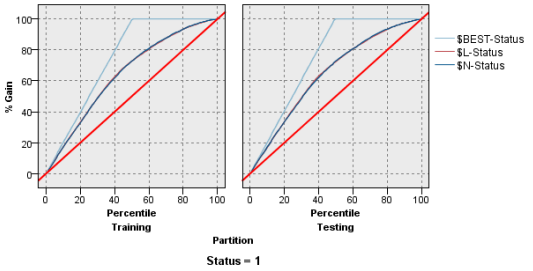
	R...	Field	Measurement	Importance	Value
✓	1	age	Continuous	Important	1.0
✓	2	storecar	Continuous	Important	1.0
✓	3	numcards	Continuous	Important	1.0
✓	4	numkids	Continuous	Important	1.0
✓	5	marital	Nominal	Important	1.0
✓	6	loans	Continuous	Important	1.0
✓	7	howpaid	Nominal	Important	0.991
	8	income	Continuous	Marginal	0.948
	9	mortgage	Nominal	Unimportant	0.442
	10	gender	Nominal	Unimportant	0.071

Selected fields: 7 Total fields available: 10

> 0.95 <= 0.95 < 0.9

Настройка модели и оценка точности

- Модели классификации, как правило, настраиваются на сбалансированной обучающей выборке. Точность модели оценивается на обучающей и контрольной выборке. Оценка точности модели на контрольной выборке позволяет судить об устойчивости модели. Модель считается устойчивой, когда ее точность на контрольной выборке близка к точности, полученной на обучающей выборке.
- При настройке нескольких видов моделей осуществляется сравнение точности и устойчивости моделей с использованием обучающей, тестовой и валидационной выборок. Для оценки используются также различные графики, характеризующие избирательность моделей, например, Gain chart. Такие графики позволяют оценить не только интегральную точность модели, но и точность на различных процентилях отбора.
- Процесс настройки и оценки качества модели может выполняться итерационно, с внесением изменений в параметры модели и/или в состав предикторов по результатам оценки качества.



Results for output field Status

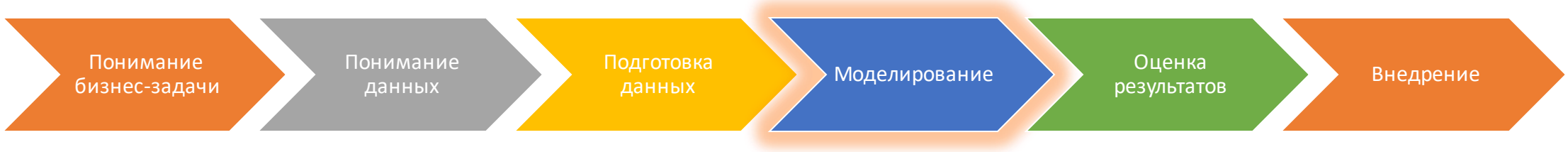
Individual Models

Comparing $L-Status with Status

'Partition'	1_Training		2_Testing	
Correct	26,334	72.87%	11,188	72.24%
Wrong	9,803	27.13%	4,300	27.76%
Total	36,137		15,488	

Comparing $N-Status with Status

'Partition'	1_Training		2_Testing	
Correct	26,329	72.86%	11,236	72.55%
Wrong	9,808	27.14%	4,252	27.45%
Total	36,137		15,488	



Оценка результатов

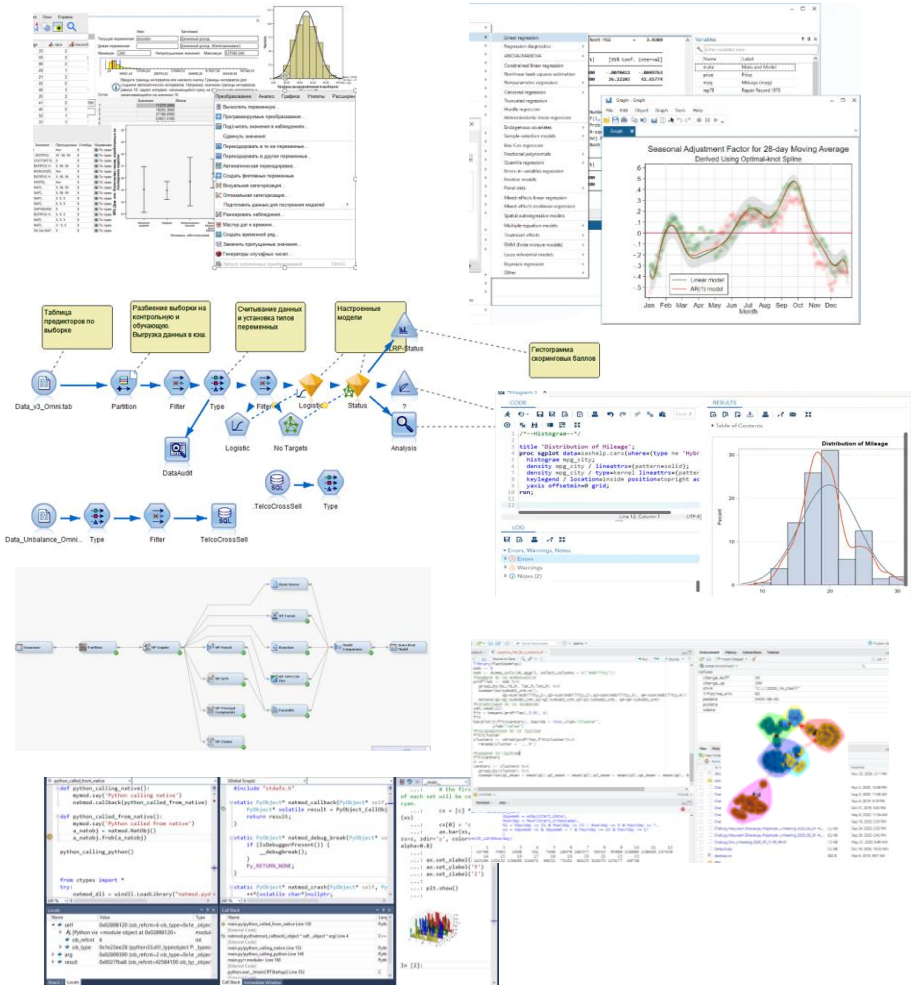
- В отличие от оценки точности модели на этапе моделирования, данный этап Data mining предполагает оценку степени соответствия полученных результатов настройки модели поставленной цели. В данном случае модель оценивается с точки зрения бизнес-результатов, достижение которых она обеспечивает.
- На данном этапе также проверяется, нет ли проблем с качеством модели, например, корректно ли построена модель, используются ли в модели только те предикторы, которые будут доступны на этапе применения (прогнозирования - скоринга).
- Определение следующих шагов: переход к внедрению (практическому использованию) модели; формулирование задач следующего проекта; возврат к этапу сбора данных и моделирования в случае необходимости повышения качества модели.

Внедрение результатов моделирования в бизнес-процессы

- Внедрение модели или моделей обычно представляет собой разработку процесса регулярного применения модели, например, разработку процесса скоринга по склонности к отклику на предложение.
- Результатом внедрения модели может быть автоматизированный процесс применения модели или регулярный расчет результатов прогнозирования с использованием инструментов используемых программных средств анализа данных.
- Этап внедрения модели включает также планирование мониторинга точности модели и потенциальных действий в случае снижения точности.

Примеры программных средств для анализа данных

- IBM SPSS
  - IBM SPSS Statistics – статистический пакет
  - IBM SPSS Modeler – платформа Data mining
- SAS
  - SAS Base – статистический пакет
  - SAS Enterprise Miner – платформа Data mining
- Stata – статистический пакет
- Statistica Data Miner – платформа Data mining
- Matlab – пакет анализа, математического моделирования
- R, Python – языки программирования с развитыми средствами и библиотекам анализа данных

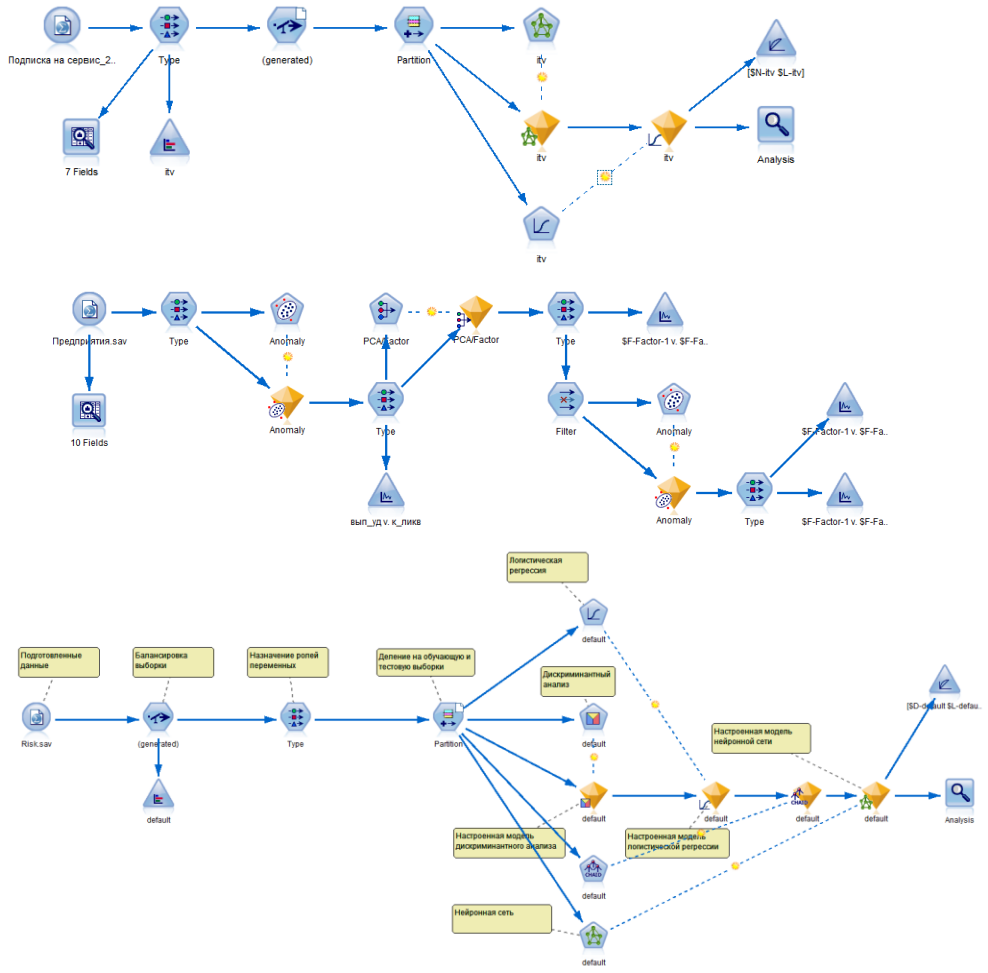


Основные классы методов моделирования и примеры алгоритмов

- Классификация
  - Дискриминантный анализ
  - Логистическая регрессия
  - Деревья решений «классические»: CHAID, C&RT, C5
  - Комбинирование деревьев решений: XGBoost, Random Forest
  - Нейронные сети
  - Байесовские сети
  - SVM
- Прогнозирование количественных переменных
  - Линейная регрессия
  - Общие линейные модели
  - Деревья решений: C&RT, CHAID
  - Метод ближайшего соседа KNN
- Классификация без учителя
  - Кластерный анализ К-средних
  - Двухэтапный кластерный анализ
  - Нейронная сеть Кохонена
- Обнаружение аномалий
  - Anomaly (на основе кластеризации)
- Ассоциативные правила
  - Apriory, Sequence, Association Rules
- Прогнозирование временных рядов
  - Экспоненциальное сглаживание
  - ARIMA
  - Нейронные сети с глубоким обучением (TCNN, RNN)

Демонстрация

- Отклик на подписку itv
- Оборот супермаркета в зависимости от характеристик месторасположения
- Обнаружение аномалий: характеристики предприятий
- Кредитные риски

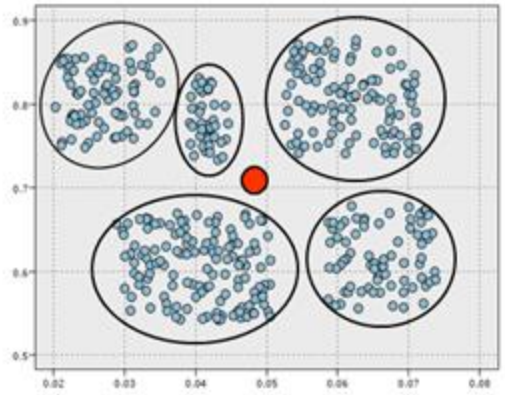


Спасибо, пожалуйста, вопросы

Spare slides

Обнаружение аномалий на основе кластерного анализа

Идея заключается в анализе не просто отклонений отдельных показателей от их средних значений или квартилей, а в анализе сочетаний показателей, выявлении кластеров в пространстве показателей и обнаружении отклонения точек от соответствующих кластеров.

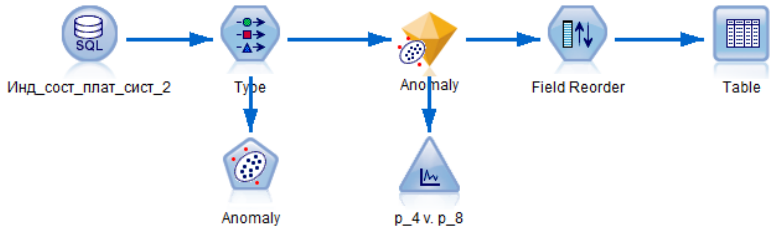


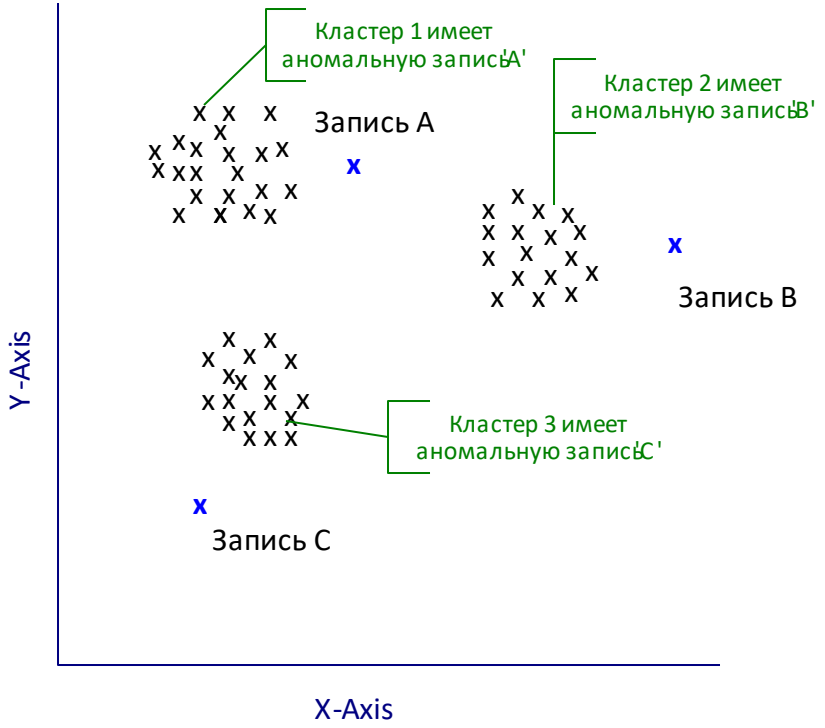
Области применения:

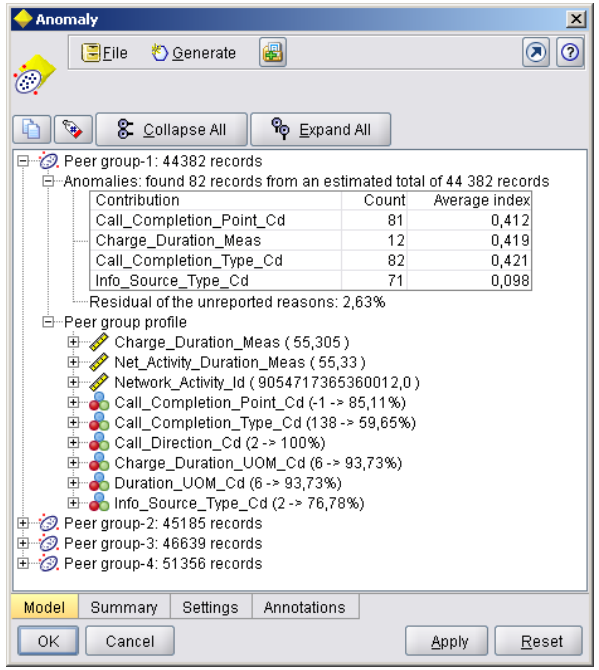
- Очистка данных от аномальных наблюдений для повышения качества настройки прогностических моделей
- Идентификация необычных наблюдений в интересах обнаружения случаев мошенничества – интерес представляют сами аномальные наблюдения

Обнаружение аномалий

- Поиск необычных наблюдений в данных
- Обнаружение аномалий на основе кластеризации
- Применение в сочетании с факторным анализом





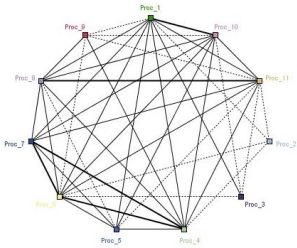


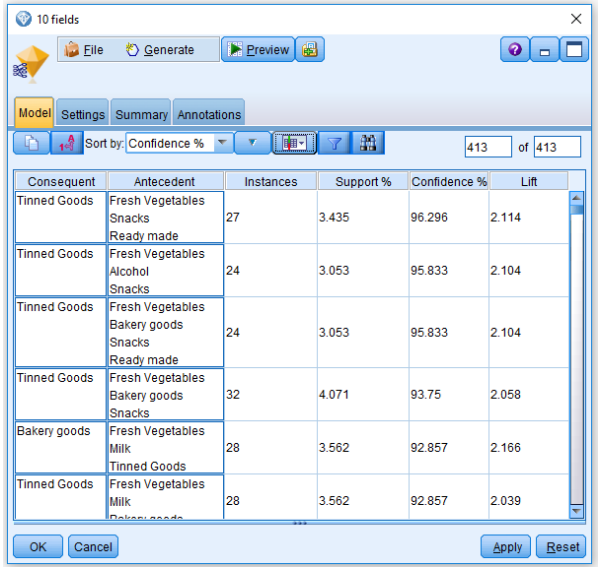
Ассоциативные правила

Алгоритмы поиска ассоциаций обнаруживают взаимосвязи между категориями, например, покупаемых товаров, которые встречаются вместе.

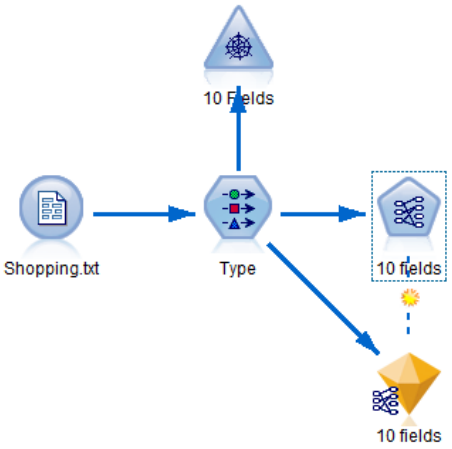
Существуют разновидности алгоритмов, которые обнаруживают последовательности, например, типичные последовательности просматриваемых веб-страниц.

Ассоциативные правила находят применение в системах рекомендаций.





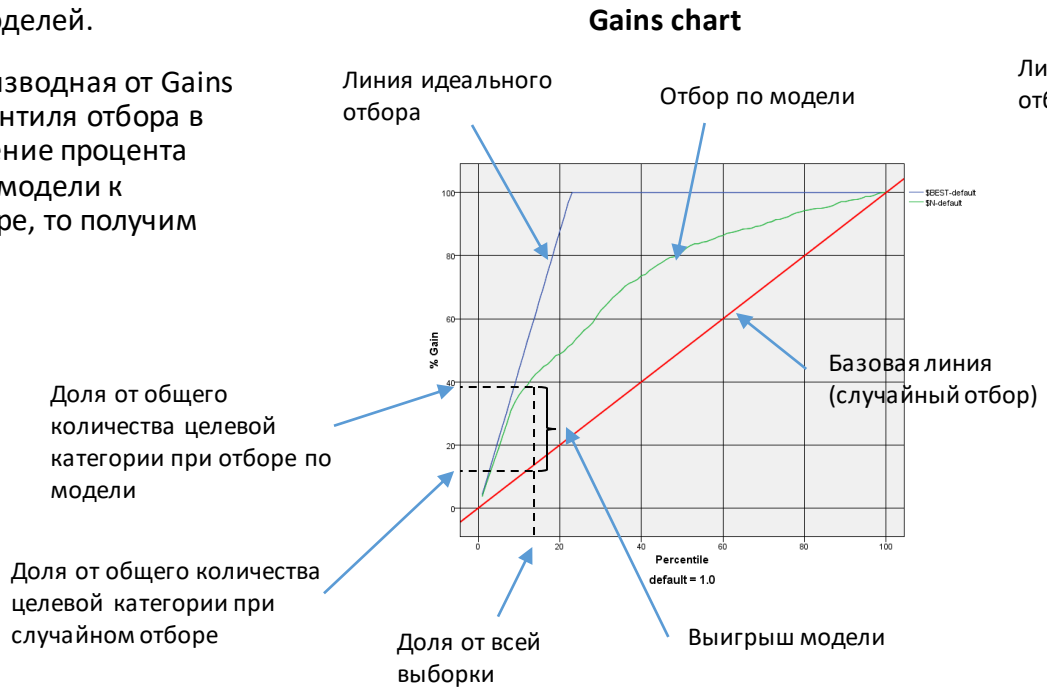
Consequent	Antecedent	Instances	Support %	Confidence %	Lift
Tinned Goods	Fresh Vegetables Snacks Ready made	27	3.435	96.296	2.114
Tinned Goods	Fresh Vegetables Alcohol Snacks	24	3.053	95.833	2.104
Tinned Goods	Fresh Vegetables Bakery goods Snacks Ready made	24	3.053	95.833	2.104
Tinned Goods	Fresh Vegetables Bakery goods Snacks	32	4.071	93.75	2.058
Bakery goods	Fresh Vegetables Milk Tinned Goods	28	3.562	92.857	2.166
Tinned Goods	Fresh Vegetables Milk	28	3.562	92.857	2.039

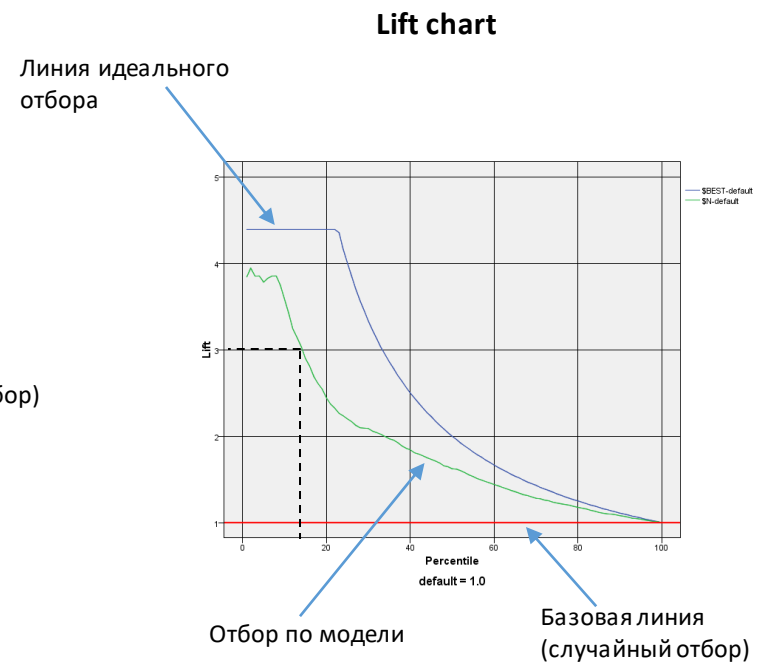


Графики выигрыша (Gain) и роста (Lift) для оценки ожидаемого эффекта моделей

Графики выигрыша (Gains chart) позволяют оценить избирательность модели при отборе ТОП N% по результатам скоринга. С помощью графиков Gains можно сравнивать характеристики различных моделей.

График роста (Lift chart) – производная от Gains chart. Если для каждого процентиля отбора в Gains chart подсчитать отношение процента отбора целевой категории по модели к проценту при случайном отборе, то получим график Lift chart.

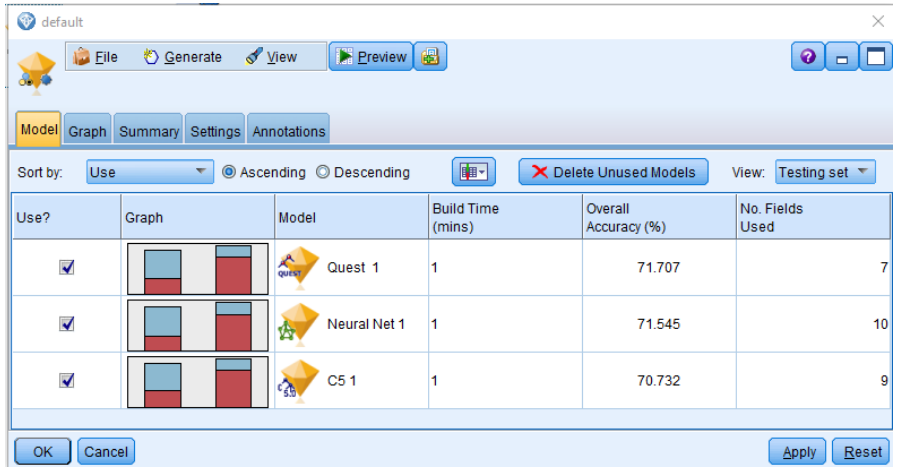




Автоматическая настройка множества моделей, комбинирование результатов

В IBM SPSS Modeler имеется возможность автоматической настройки сразу множества моделей. По результатам настройки можно выбрать модель по какому-либо критерию или использовать все настроенные модели в целях повышения надежности решения путем комбинации их результатов.

- Auto Classifier – моделирование категориальной целевой переменной
- Auto Numeric – моделирование количественной целевой переменной
- Auto Cluster – настройка нескольких различных моделей кластеризации (К-средних, двухэтапный кластерный анализ, нейронная сеть Кохонена)

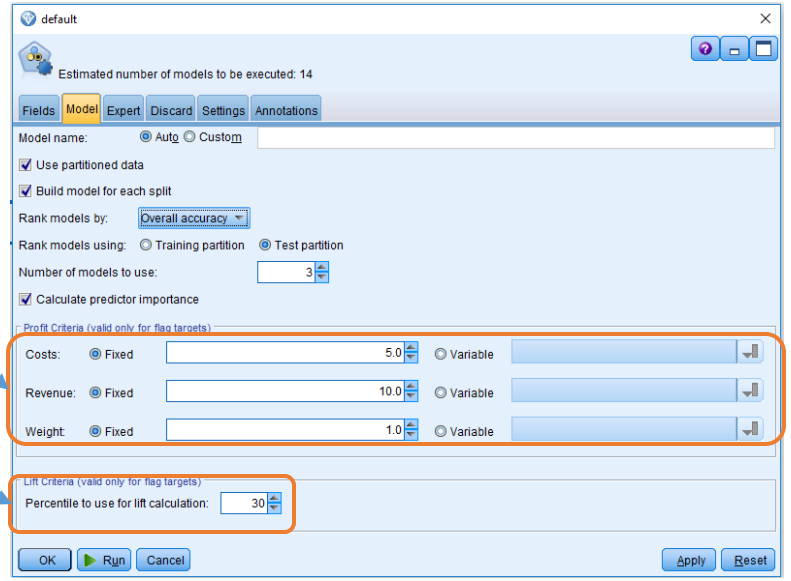


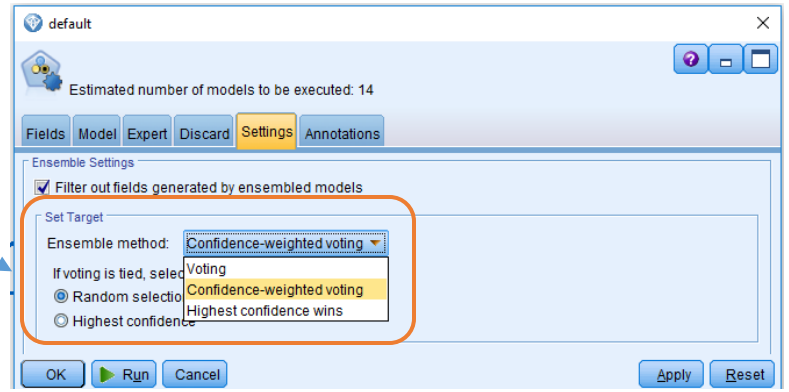
Автоматическая настройка моделей: Auto Classifier

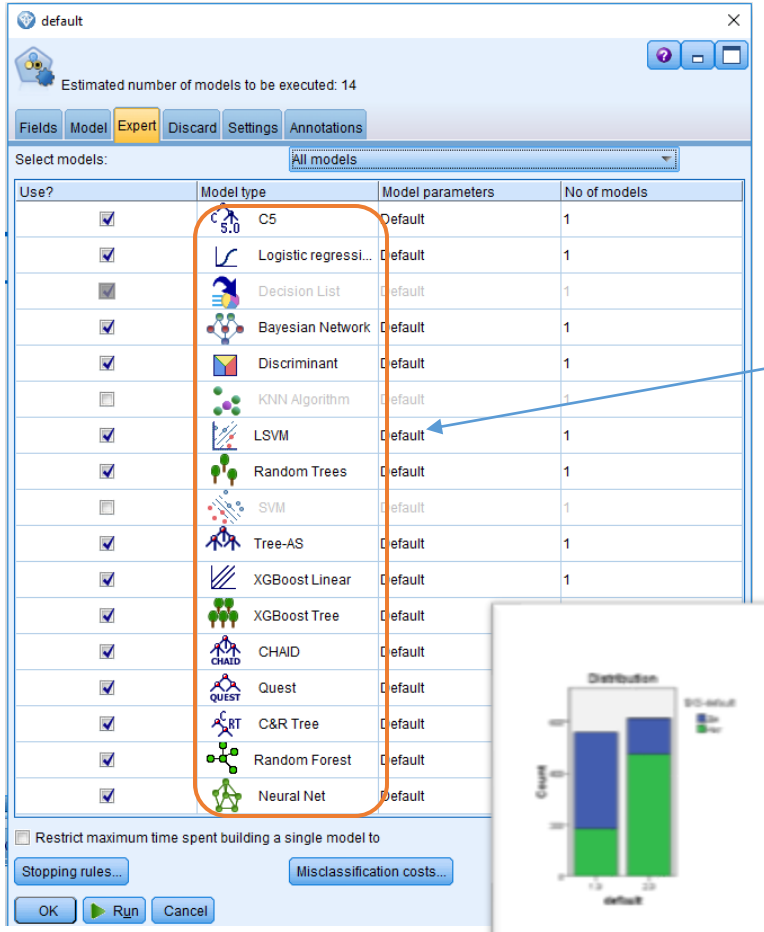
Возможность взвешивания наблюдений, задания стоимости ошибок и прибыли

Выбор процентиля для сравнения моделей по показателю Lift

Критерии совместного использования моделей, различные варианты голосования







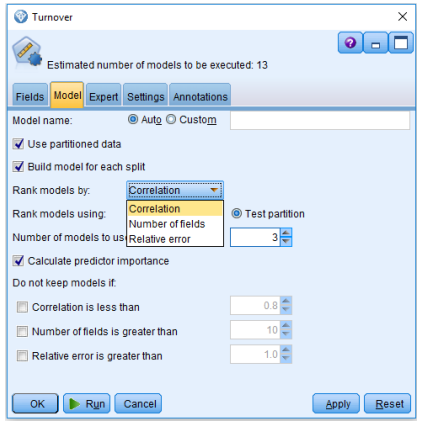
Auto Classifier

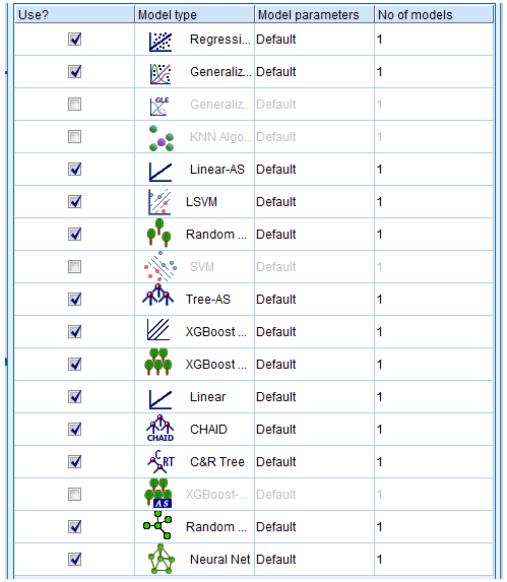
Настраиваемые модели

Автоматическая настройка моделей: Auto Numeric

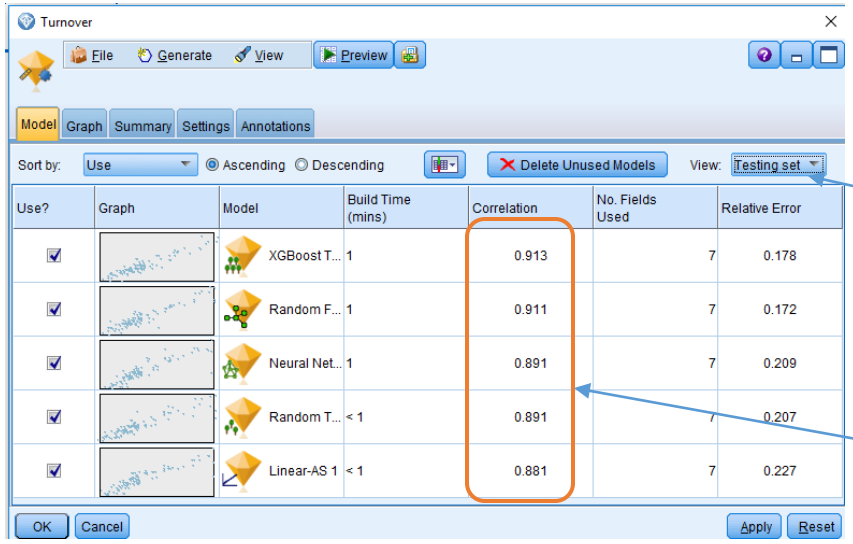
Автоматическая настройка моделей с количественной целевой переменной.

Одновременное применение моделей или выбор наилучшей модели.



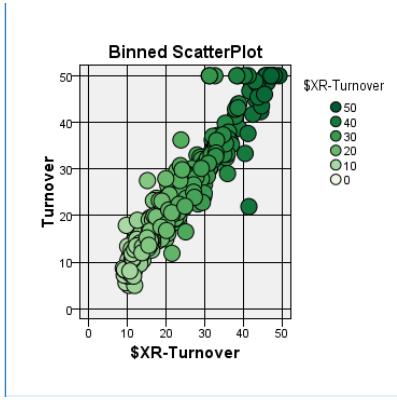


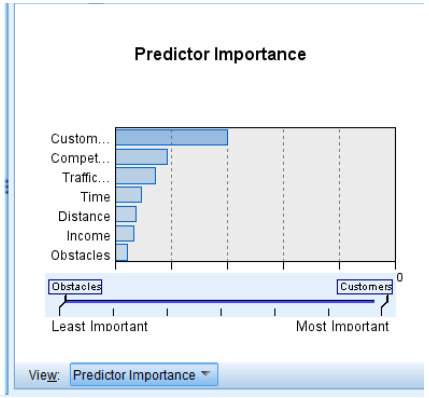
Use?	Model type	Model parameters	No of models
☑	Regressi...	Default	1
☑	Generaliz...	Default	1
☐	Generaliz...	Default	1
☐	KNN Algo...	Default	1
☑	Linear-AS	Default	1
☑	LSVM	Default	1
☑	Random ...	Default	1
☐	SVM	Default	1
☑	Tree-AS	Default	1
☑	XGBoost ...	Default	1
☑	XGBoost ...	Default	1
☑	Linear	Default	1
☑	CHAID	Default	1
☑	C&R Tree	Default	1
☐	XGBoost-...	Default	1
☑	Random ...	Default	1
☑	Neural Net	Default	1



Обучающая/контрольная выборки

Корреляция прогноза с фактическими значениями целевой переменной





Автоматизированная подготовка данных для построения моделей

Автоматическая или интерактивная подготовка данных к анализу (Преобразования -> Подготовить данные для построения моделей)

- Подготовка дат
- Исключение полей
- Настройка шкал измерения
- Обработка выбросов и пропущенных значений
- Преобразования – стандартизация
- Преобразование категориальных переменных и категоризация количественных
- Отбор предикторов