Программа учебной дисциплины «Глубинный анализ данных и текстов на базе IBM SPSS Modeler»

Утверждена
Академическим советом ООП
Протокол № от «26» июня 2018 г.

Автор	Бекларян Армен Левонович
Число кредитов	6
Контактная работа (час.)	72
Самостоятельная работа (час.)	156
Курс	2
Формат изучения дисциплины	без использования онлайн курса

ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целью освоения учебной дисциплины является формирование у студентов комплекса теоретических знаний, методологических основ и практических навыков в области повышение эффективности политики удержания клиентов, стимуляция кросс-продаж и повторных покупок, сегментация клиентов, минимизация кредитных рисков, а также обнаружение и предотвращение мошенничества с использованием инструмента IBM SPSS Modeler. Подобные компетенции дают возможность анализировать тенденции, закономерности и взаимосвязи в структурированных и неструктурированных данных, прогнозировать на основе этого анализа будущие события и действовать для достижения желаемых результатов. Программная платформа IBM SPSS Modeler – это мощная платформа прогнозной аналитики, позволяющая извлекать из данных беспрецедентные объемы ценной информации, строить на ее основе прогнозы и принимать эффективные решения на всех уровнях управления.

Задачи курса:

– Формирование теоретических и методологических основ в области обнаружении в данных устойчивых закономерностей, которые могут быть использованы для принятия решений и управления взаимоотношениями с клиентами.

– Формирование практических навыков использования интерфейса визуального программирования для построения потоков обработки данных и их моделирования.

– Формирование навыков построения прогнозных моделей, использующих деловые знания и опыт, и внедрять их в деловые операции для усовершенствования процесса принятия решений.

В результате освоения дисциплины студент должен:

знать:

– алгоритмы и методы анализа данных, в том числе анализ текста, анализ сущностей, управление решениями и их оптимизацию, что позволяет получать знания в режиме реального времени;

– весь технологический процесс непрерывного поиска ценной информации в различных источниках данных с оперативным внедрением найденных закономерностей в практику бизнеса.

уметь:

– извлекать из различных источников данные, необходимые для решения задач бизнес анализа в соответствии с этапами Межотраслевого стандарта Data Mining (CRISP-DM), использовать интерфейс визуального программирования для построения потоков обработки данных и моделирования;

– осуществлять выбор локального, облачного или гибридного вариантов развертывания для получения прогнозной аналитики с использованием встроенных служб, интеграции бизнес-аналитики и формирования простой отчетности.

владеть:

– навыками анализа многомерных данных с использованием методов деревьев решений, кластеризации и ассоциативных правил с применением программного продукта IBM SPSS Modeler;

– навыками быстрой разработки точных моделей прогнозирования и применять прогнозную аналитику на уровне отдельных пользователей, групп, систем и всего предприятия.

Изучение дисциплины «Глубинный анализ данных и текстов на базе IBM SPSS Modeler» базируется на следующих дисциплинах:

– Анализ неструктурированной информации;
– Системный анализ и проектирование;
– Прогностическая аналитика;
– Системы поддержки принятия решений.

Для освоения учебной дисциплины, студенты должны знать концептуальные основы архитектуры предприятия, основные классы информационных систем управления бизнесом, лучшие практики и современные стандарты в сфере информационных технологий.

Также студенты должны владеть методами проектирования информационных систем, уметь систематизировать и обобщать информацию, разрабатывать конкретные предложения по результатам исследований, готовить справочно-аналитические материалы для принятия управленческих решений в сфере информационных технологий.

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин:

– Системы имитационного моделирования;
– Прогнозирование бизнес-результатов деятельности предприятия на основе предиктивного моделирования с использованием IBM Watson;
– Информационные технологии в анализе инвестиционных проектов.

I. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Введение в IBM SPSS Modeler

Обзор создания потока. Построение потоков данных. Работа со узлами. Работа с потоками. Описания потока. Выполнение потоков. Работа с моделями. Добавление комментариев и аннотаций к узлам и потокам. Сохранение потоков данных. Загрузка файлов. Отображение потоков данных.

Средство просмотра. Показ и скрытие результатов. Перемещение, копирование и удаление результатов. Изменение исходного выравнивания. Изменение выравнивания элементов вывода. Схема вывода. Добавление элементов в средстве просмотра. Поиск и замена информации в средстве просмотра. Обзор пропущенных значений. Обработка пропущенных значений.

Тема 2. Анализ данных с использованием IBM SPSS Modeler

Параметры алгоритма для узла автоматического моделирования. Правила остановки для узла автоматического моделирования. Узел автоклассификации. Опции моделей узла автоклассификации. Дополнительные опции узла автоклассификации. Стоимости ошибочной классификации. Опции отклонения узла автоклассификации. Опции параметров узла автоклассификации. Узел автонумерации. Опции моделей узла автонумерации. Опции эксперта узла автонумерации. Опции параметров узла автонумерации. Узел автокластеризация. Опции модели узла автоматической кластеризации. Опции эксперта узла автокластеризация. Опции отбрасывания узла автокластеризации. Слепки автоматизированных моделей. Генерирование узлов и моделей. Генерирование диаграмм оценки. Графики оценки.

Узел Кохонена. Опции моделей узла Кохонена. Дополнительные опции узлов Кохонена. Слепки моделей Кохонена. Сводка модели Кохонена. Узел k-средних. Опции моделей узла k-средних. Опции эксперта узла k-средних. Слепки моделей k-средних. Сводка моделей k-средних. Узел двухшаговой кластеризации. Опции модели узла двухшаговой кластеризации. Слепки двухшаговых моделей кластеров. Сводка двухшаговой модели. Узел кластера TwoStep-AS. Кластерный анализ Twostep-AS. Слепки моделей кластеров TwoStep-AS. Параметры слепков моделей кластеров TwoStep-AS. Средство просмотра кластеров. Построение диаграмм на основе моделей кластеров.

Тема 3. Предиктивный анализ с использованием IBM SPSS Modeler

Модели деревьев решений. Интерактивный построитель деревьев Слепок модели деревья решений. Средство просмотра деревья решений. Модель нейросетей. Использование нейронных сетей совместно с унаследованными потоками. Целевые показатели. Основные параметры. Правила остановки. Ансамбли. Дополнительные опции модели. Сводка для модели. Важность предикторов. Предсказанные против наблюденных. Классификация.

Линейные модели. Линейные-AS модели. Логистический узел. Опции моделей узла логистической регрессии. Опции сходимости логистической регрессии. Расширенный вывод для логистической регрессии. Опции шагового отбора логистической регрессии. Узел PCA/фактора. Опции моделей узла PCA/факторной модели. Узел дискриминанта. Узел обобщенной линейной модели. Расширенный вывод для обобщенных линейных. Обобщенные линейные смешанные модели. Узел байесовская сеть. Опции модели узла байесовской сети. Дополнительные опции узла байесовской сети. Слепки моделей байесовской сети. Параметры модели байесовской сети. Сводка моделей байесовской сети.

Тема 4. Анализ текстов с использованием IBM SPSS Modeler

Чтение в исходном тексте. Принципы работы извлечения. Принципы работы категоризации. Узел списка файлов. Использование узла список файлов для исследования текстовых данных. Узел Веб-фид. Использование узла веб-фидов для исследования текстовых данных.

Режим моделирования Text Mining. Добавление расположенного выше узла выборки для экономии времени. Использование узла Text Mining в потоке. Слепок Text Mining: модель понятий. Использование слепков модели понятий в потоке. Слепок Text Mining: модель категорий. Использование слепков модели категорий в потоке.

III. ОЦЕНИВАНИЕ

Формами текущего контроля являются контрольная работа и домашнее задание. Каждая из форм текущего контроля оценивается по 10-балльной шкале. Общая оценка за текущий контроль (по 10-балльной шкале) рассчитывается по формуле:

$$O_{текущий} = 0{,}4 \cdot O_{кр} + 0{,}6 \cdot O_{дз},$$

где $O_{кр}$ – оценка за контрольную работу;

$O_{дз}$ – оценка за домашнее задание.

При определении накопленной оценки (по 10-балльной шкале) самостоятельная внеаудиторная работа не оцениваются. Поэтому накопленная оценка формируется из оценки за текущий контроль и оценки за работу на аудиторных занятиях, и рассчитывается по формуле:

$$O_{накопленная} = 0{,}8 \cdot O_{текущий} + 0{,}2 \cdot O_{ауд} + 0{,}0 \cdot O_{сам.работа},$$

где $O_{текущий}$ – оценка за текущий контроль;

$O_{ауд}$ – оценка за аудиторную работу;

$O_{сам.работа}$ – оценка за самостоятельную работу.

Оценка за аудиторную работу выставляется на основе пропорции посещаемости студента к общему числу проведенных занятий, исходя из максимума в 10 баллов.

Результирующая оценка (выставляется в диплом) формируется на основе итоговой оценки за экзамен (по 10-балльной шкале) и накопленной оценки. Результирующая оценка рассчитывается по формуле:

$$O_{результ} = 0{,}4 \cdot O_{экз} + 0{,}6 \cdot O_{накопленная},$$

где $O_{экз}$ – оценка за итоговый контроль (экзамен);

$O_{накопленная}$ – накопленная оценка.

$$O_{экз} = 0{,}5 \cdot O_{экз1} + 0{,}5 \cdot O_{экз2},$$

где $O_{экз1}$ – оценка за тестовую часть экзамена;

$O_{экз2}$ – оценка за практическую часть экзамена.

Оценка за тестовую часть экзамена представляет собой сумму баллов за каждый вопрос, при этом вопросы дают вклад либо 1 балл, в случае полностью верного данного ответа, либо 0 баллов, в противном случае. Для вопросов с множественным выбором правильным считается тот ответ, в рамках которого выбраны все правильные варианты ответа и только они.

Оценка практической части экзамена эквивалентны правилам оценивания контрольной работы.

При формировании результирующей оценки на основе весовых коэффициентов применяется арифметическое округление до целого числа. В случае точного равенства дробной части пяти десятым округление применяется в большую сторону.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

4.1 Содержание заданий контроля

Выполнение домашнего задания предусматривает построение моделей анализа информации и текстов, выявление регулярных выражений, построение аналитических срезов и фильтров, выделение корреляций между срезами, отображение взаимосвязей и визуализацию итогов анализа в системе IBM SPSS Modeler.

Контрольная работа формируется на основе практических занятий.

Экзаменационная работа состоит из двух частей: тест и практическая часть. Тест представляет из себя 10 вопросов закрытого типа, практическая часть – проведение анализа данных заданного типа, на основе видов молей, пройденных после проведения контрольной работы.

4.2 Пример задания контрольной работы и практической части экзамена

Цель – построить модель, определяющую относится ли человек к группе риска сердечного приступа.

Необходимые шаги:

1. Используйте файл с данными о пациентах, содержащий информацию о состоянии здоровья и поведенческих особенностях.
2. Определите, какие поля будут использоваться в модели. В случае необходимости проведите модификацию значений, а также постройте новые вычислимые поля.
3. Выберите один или несколько типов анализа. Обоснуйте свой выбор.
4. Постройте модель, опишите результат, а также примените модель к проверочной выборке.

4.3 Пример вопросов тестовой части экзамена

1. Продавец проводит анализ покупок своих клиентов с целью выявления самых популярных комбинаций продуктов. Это пример проведения:

   А. Классификации
   Б. Сегментации
   В. Выявление взаимосвязи

2. Выберите верное утверждение:

   А. Quest алгоритм всегда строит бинарное дерево
   Б. Quest алгоритм использует суррогатные поля для классификации записи с пропущенными значениями по значимому полу
   В. Quest алгоритм использует одинаковый тест значимости для порядковых и непрерывных полей

4.4 Методические указания студентам по подготовке к контрольной работе

В рамках контрольной работы студентам будет предложено решить практический кейс по анализу данных компании и построению модели заданного типа (предиктивная, регрессионная, кластерная и др.), ограничиваясь видами моделей, пройденными по курсу к моменту проведения контрольной работы. Предлагаемый кейс аналогичен разбираемому на практических занятиях.

Отчетные материалы:

1. Исходные данные (csv, xml, json, xlsx и др.)
2. Проект IBM SPSS Modeler
3. Результирующие данные (результат предиктивного анализа, регрессии, кластеризации и др.)
4. Сопроводительная документация (описание модели, обоснование проделанных шагов)

Оценивание:

Построение корректной модели в IBM SPSS Modeler дает возможность получения оценки 4 и является обязательным минимумом при выполнении задания.

Применение построенной модели к проверочной выборке добавляет максимум 2 балла.

Приведение описания построенной модели, ее анализ (точность, устойчивость и пр.), а также обоснование шагов построения модели добавляет максимум 4 балла.

4.5 Методические указания студентам по подготовке домашнего задания

В рамках домашнего задания студентам необходимо подготовить аналитическую отчетность по выбранной предметной области с использованием IBM SPSS Modeler.

1. Источник данных
Источник данных студент выбирает самостоятельно также, как и предметную область. Среди обязательных требований наличие не менее 10000 записей, либо символов в случае анализа текста.

2. Модификация данных и первичный анализ
Модификация исходных данных в системе IBM SPSS Modeler, а также предварительный анализ в форме отчета верхнего уровня формируется в системе Microsoft Power BI.

3. Внедрение вычислимых полей
Вычислимые поля могут представлять из себя как промежуточные вычисления в рамках статистического анализа (п.4), так и введенные KPI или иные меры на исходных данных.

4. Проведение статистического и иного анализа
Статистические анализ проводится с целью выявления скрытых зависимостей или для подтверждения гипотез, сформированных по итогам первичного анализа (п.2), в частности, допустимо проведение регрессионного анализа. Можно также использовать интеграционный модуль с системами статистического анализа R или SPSS Statistics.

5. Подготовка пояснительной записки
Формирование отчета по результатам проделанной работы, включая, но не ограничиваясь, описанием исходных данных, методами и итогами статического анализа, выводами.

Отчетные материалы:
1. Исходные данные (csv, xml, json, xlsx и др.)
2. Проект IBM SPSS Modeler
3. Скрипты статистического анализа
4. Сопроводительная документация

Оценивание:
Выполнение пп.1 и 2 дает возможность получения оценки 4 и является обязательным минимумом при выполнении задания.
Выполнение п.3 добавляет максимум 1 балл.
Выполнение п.4 добавляет максимум 3 балла.
Выполнение п.5 добавляет максимум 2 балла.

V. РЕСУРСЫ

5.1 Основная литература

1. Кудрявцев В.Б. Интеллектуальные системы: учебник и практикум для бакалавриата и магистратуры / В.Б. Кудрявцев, Э.Э. Гасанов, А.С. Подколзин. – 2-

е изд., испр. и доп. – М.: Издательство Юрайт, 2017. – 219 с. Режим доступа: www.biblio-online.ru/book/1DAA117E-A40C-4F22-B6EA-642C255D29CB.
2. Миркин Б.Г. Введение в анализ данных: учебник и практикум для вузов / Б.Г. Миркин. – М.: Юрайт, 2015. – 174 с.
3. Подкорытова О.А. Анализ временных рядов: учеб. пособие для бакалавриата и магистратуры / О.А. Подкорытова, М.В. Соколов. – 2-е изд., перераб. и доп. – М.: Издательство Юрайт, 2017. – 267 с. Режим доступа: www.biblio-online.ru/book/634E8D89-2B9B-483C-8985-18666AD3B04A.
4. Станкевич Л.А. Интеллектуальные системы и технологии: учебник и практикум для бакалавриата и магистратуры / Л.А. Станкевич. – М.: Издательство Юрайт, 2017. – 397 с. Режим доступа: www.biblio-online.ru/book/962CB15C-CEA9-4134-B1B1-4DC4A85783AA.

5.2 Дополнительная литература

1. Бринк Х., Ричардс Д., Феверолф М. Машинное обучение. Питер, 2017.
2. Вьюгин В.В. Математические основы теории машинного обучения и прогнозирования. 2-е издание, исправленное и дополненное. – М.: МЦНМО, 2018. – 384 с.
3. Груздев А. Прогнозное моделирование в IBM SPSS Statistics и R. Метод деревьев решений. – М.: ДМК Пресс, 2016 – 278 с.
4. Груздев А. Прогнозное моделирование в IBM SPSS Statistics, R и Python. Метод деревьев решений и случайный лес. – М.: ДМК Пресс, 2017 – 634 с.
5. Моосмюллер Г., Ребик Н.Н. Маркетинговые исследования с SPSS: Учебное пособие. – М.: ИНФРА-М, 2018.
6. Орлова И. Многомерный статистический анализ в экономических задачах. Компьютерное моделирование в SPSS. – М.: Вузовский учебник, 2017. – 310 с.
7. Marvin L. Decision Trees and Applications with IBM SPSS Modeler. CreateSpace Independent Publishing Platform, 2016.
8. McCormick K. et al. IBM SPSS Modeler Cookbook. Packt Publishing, 2013.
9. McCormick K., Salcedo J. IBM SPSS Modeler 19 Essentials. Packt Publishing, 2017.
10. Rosius W. Introduction to R in IBM SPSS Modeler. IBM Redbooks, 2016.
11. Wendler T., Gröttrup S. Data Mining with SPSS Modeler. Springer Nature, 2016.

5.3 Справочники, словари, энциклопедии

1. Руководство пользователя IBM SPSS Modeler 18. [URL]: https://www.ibm.com/support/knowledgecenter/ru/SS3RA7_18.0.0/modeler_mainhelp_client_ddita/modeler_mainhelp_client_ddita-gentopic2.html.
2. IBM SPSS Modeler documentation. [URL]: http://www.ibm.com/support/knowledgecenter/SS3RA7_18.0.0/clementine/knowledge_center/product_landing.html
3. SPSS Modeler 18.0 Documentation. [URL]: http://www-01.ibm.com/support/docview.wss?uid=swg27046871

5.4 Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Microsoft Windows 7 Professional или более новая версия	Из внутренней сети университета (договор)
2.	Microsoft Office Professional Plus 2013 или более новая версия	Из внутренней сети университета (договор)

3.	Microsoft SQL Server 2014 Enterprise Edition или более новая версия	Из внутренней сети университета (договор)
4.	Microsoft Power BI	Свободно распространяемое ПО
5.	IBM SPSS Modeler 18 Premium или более новая версия	Из внутренней сети университета (договор). Дистрибутив предоставляется в рамках программы IBM Academic Initiative
6.	IBM SPSS Modeler R Essentials версия аналогична версии IBM SPSS Modeler	Из внутренней сети университета (договор). Дистрибутив предоставляется в рамках программы IBM Academic Initiative
7.	R 3.1.2 или более новая версия	Из внутренней сети университета (договор)
8.	RStudio	Из внутренней сети университета (договор)
9.	Anaconda 3 x64	Из внутренней сети университета (договор)
10.	Faronics Insight	Из внутренней сети университета (договор)

5.5 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
	Профессиональные базы данных, информационно-справочные системы	
1.	Электронно-библиотечная система Юрайт	URL: https://biblio-online.ru/
	Интернет-ресурсы (электронные образовательные ресурсы)	
1.	Программный хаб IBM	URL: https://ibm.onthehub.com
2.	Образовательный портал IBM	URL: https://www.ibm.com/developerworks/ru/

5.6 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

– ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);

– мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены ПЭВМ, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ, а также с установленным требуемым программным обеспечением, в количестве одна единица на каждого слушателя дисциплины.