Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»

УТВЕРЖДАЮ
Директор Института магистратуры
Иванова Е.А.
«01» июня 2023г.

Рабочая программа дисциплины
Методы анализа больших данных (Big Data)

Направление 38.04.01 Экономика
магистерская программа 38.04.01.09 "Финансовый аналитик"

Для набора 2023 года

Квалификация
магистр

КАФЕДРА **Статистики, эконометрики и оценки рисков**

Распределение часов дисциплины по семестрам

Семестр (<Курс>.<Семестр на курсе>)	**1 (1.1)**		Итого	
Недель	15 2/6			
Вид занятий	УП	РП	УП	РП
Лекции	16	16	16	16
Лабораторные	10	10	10	10
Практические	10	10	10	10
Итого ауд.	36	36	36	36
Контактная работа	36	36	36	36
Сам. работа	36	36	36	36
Итого	72	72	72	72

ОСНОВАНИЕ

Учебный план утвержден учёным советом вуза от 28.03.2023 протокол № 9.

Программу составил(и): к.э.н., доцент, Кракашова О.А.

Зав. кафедрой: д.э.н., профессор, Ниворожкина Л.И.

Методическим советом направления: д.э.н., профессор, Ниворожкина Л.И.

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

1.1	Сформировать у студентов системное представление о технологиях многомерного анализа данных, интеллектуального анализа данных (Data Mining), их применении и инструментах, изучить основные методы прикладного анализа данных, развить навыки исследования различных процессов с использованием современных информационно-коммуникационных технологий, практического применения методов многомерного анализа и Data Mining для решения различных научных и технических задач в экономике и бизнесе.

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

ПК-4:Способен формировать информационную базу бизнес-анализа, оценивать текущее состояние организации (объекта исследования) выявлять, и оценивать несоответствия между параметрами ее текущего и будущего состояний

В результате освоения дисциплины обучающийся должен:

Знать:

базовые понятия и современные информационные технологии Big Data, а также языки визуального моделирования применяемые для целей бизнес-анализа (соотнесено с индикатором ПК-4.1).

Уметь:

анализировать большие массивы данных, характеризующие различные социально-экономические процессы, внутренние (внешние) факторы и условия, влияющие на деятельность организации; применять информационные технологии анализа Big Data в объеме, необходимом для целей бизнес-анализа (соотнесено с индикатором ПК-4.2).

Владеть:

современными технологиями создания и анализа больших данных, а также навыками применения информационных технологии в объеме, необходимом для целей бизнес-анализа (соотнесено с индикатором ПК-4.3).

3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Код занятия	Наименование разделов и тем /вид занятия/	Семестр / Курс	Часов	Компетен-ции	Литература
	Раздел 1. Сбор, хранение и анализ больших данных				
1.1	Обзор Big Data. Методы и средства. Используемые программы. Технологии хранения больших данных. /Лек/	1	2	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
1.2	Введение в R. R: начало работы, ввод данных и работа с большими массивами данных. Работа с электронными таблицами больших данных в Calc. /Лаб/	1	2	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
1.3	Методы и средства анализа Big Data. Используемые программы. Технологии хранения больших данных. Введение в R (ввод данных, работа с большими массивами данных). /Ср/	1	4	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
1.4	Определение источника больших данных. Исследование источника данных. Хранилище данных. /Лек/	1	2	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.4 Л2.5 Л2.6 Л2.7 Л2.8 Л2.9 Л2.10 Л2.12
1.5	R: визуализация Big Data, фиктивные переменные, прогнозы, проверка гипотез и ловушка дамми-переменных. Большие данные и скрипты. Проблемы регрессионного анализа больших данных (мультиколлинеарность и гетероскедастичность) в R. /Лаб/	1	4	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
1.6	Источник и хранилище больших данных. Большие данные в R: постороение графиков и прогнозов, доверительных и предиктивных интервалов. /Ср/	1	4	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12

1.7	Процесс анализа больших данных. Технологии анализа больших данных. Научные проблемы в области больших данных. /Лек/	1	2	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
1.8	Большие данные и скрипты. Проблемы регрессионного анализа больших данных (мультиколлинеарность и гетероскедастичность) в R. /Пр/	1	2	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
1.9	Процесс анализа больших данных. Технологии анализа больших данных. Научные проблемы в области больших данных. /Ср/	1	2	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
	Раздел 2. Методы и модели анализа больших данных				
2.1	Прогнозирование и предвидение в социально-политических и медиа процессах. Методы прогнозирования, использующие большие данные. /Лек/	1	2	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
2.2	Прогнозное моделирование: работа с регрессионными моделями больших данных в Calc и R. /Пр/	1	2	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
2.3	Методы прогнозирования в социально-экономических процессах. Модели прогнозирования: нейронные сети. /Ср/	1	4	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
2.4	OLAP-системы. /Лек/	1	2	ПК-4	Л1.1 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8
2.5	R: даты и временные ряды, загрузка больших данных и тесты на автокорреляцию, качественные переменные, предельные эффекты и ROC кривая. Экспоненциальное сглаживание Холта с корректировкой тренда в моделях временных рядов больших данных в Calc и R. /Лаб/	1	2	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
2.6	OLAP-системы. /Ср/	1	4	ПК-4	Л1.1 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8
2.7	Интеллектуальный анализ данных (Data Mining). /Лек/	1	2	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
2.8	Эконометрическое исследование на больших данных Российского мониторинга экономического положения и здоровья населения (РМЭЗ, он же RLMS). /Пр/	1	4	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.11 Л2.12
2.9	Эконометрическое исследование на больших данных Российского мониторинга экономического положения и здоровья населения (РМЭЗ, он же RLMS). /Ср/	1	12	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.11 Л2.12
2.10	Задачи и методы интеллектуального анализа больших данных. /Лек/	1	2	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
2.11	Кластерный анализ на больших данных. Анализ потребительской корзины. Использование метода k-средних для сегментирования клиентской базы. Сетевые графы и определение сообществ. /Пр/	1	2	ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
2.12	Задачи и методы интеллектуального анализа данных. Кластерный анализ на больших данных. /Ср/	1	2	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
2.13	Инструменты Data Mining. /Лек/	1	2	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12

2.14	Примеры анализа временных рядов больших данных в R.Определение выбросов. /Лаб/	1	2	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
2.15	Инструменты Data Mining. /Ср/	1	4	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.12
2.16	/Зачёт/	1	0	ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8Л2.1 Л2.2 Л2.3 Л2.4 Л2.5 Л2.6 Л2.7 Л2.8 Л2.9 Л2.10 Л2.11 Л2.12

4. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Структура и содержание фонда оценочных средств для проведения текущей и промежуточной аттестации представлены в Приложении 1 к рабочей программе дисциплины.

5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

5.1. Основная литература

	Авторы, составители	Заглавие	Издательство, год	Колич-во
Л1.1	Ратникова Т. А., Фурманов К. К.	Анализ панельных данных и данных о длительности состояний: учеб. пособие	М.: Издат. дом Высш. шк. экономики, 2014	20
Л1.2	Ниворожкина Л. И.	Статистические методы анализа данных: учеб.	М.: РИО, 2016	108
Л1.3	Герасимов А. Н., Громов Е. И., Скрипниченко Ю. С.	Эконометрика: продвинутый уровень: учебное пособие	Ставрополь: Ставропольский государственный аграрный университет (СтГАУ), 2016	https://biblioclub.ru/index.php?page=book&id=484978 неограниченный доступ для зарегистрированных пользователей
Л1.4	Неделько, В. М.	Основы статистических методов машинного обучения: учебное пособие	Новосибирск: Новосибирский государственный технический университет, 2010	http://www.iprbookshop.ru/45418.html неограниченный доступ для зарегистрированных пользователей
Л1.5	Лемешко, Б. Ю., Лемешко, С. Б., Постовалов, С. Н., Чимитова, Е. В.	Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход: монография	Новосибирск: Новосибирский государственный технический университет, 2011	http://www.iprbookshop.ru/47719.html неограниченный доступ для зарегистрированных пользователей
Л1.6	Гончарова, Н. Д., Терехова, Ю. С.	Анализ и моделирование статистических рядов: учебное пособие	Новосибирск: Сибирский государственный университет телекоммуникаций и информатики, 2016	http://www.iprbookshop.ru/69536.html неограниченный доступ для зарегистрированных пользователей
Л1.7	Дубина, И. Н.	Математико-статистические методы и инструменты в эмпирических социально-экономических исследованиях: учебное пособие	Саратов: Вузовское образование, 2018	http://www.iprbookshop.ru/76234.html неограниченный доступ для зарегистрированных пользователей
Л1.8	Брусенцев, А. Г.	Анализ данных и процессов. Ч.1. Методы статистического анализа данных: учебное пособие	Белгород: Белгородский государственный технологический университет им. В.Г. Шухова, ЭБС АСВ, 2017	http://www.iprbookshop.ru/92237.html неограниченный доступ для зарегистрированных пользователей

5.2. Дополнительная литература

	Авторы, составители	Заглавие	Издательство, год	Колич-во

	Авторы, составители	Заглавие	Издательство, год	Колич-во
Л2.1	Пересецкий А. А.	Эконометрические методы в дистанционном анализе деятельности российских банков	М.: Издат. дом Высш. шк. экономики, 2012	20
Л2.2	Арженовский С. В.	Эконометрика финансовых рынков: метод. указания по изучению дисциплины	Ростов н/Д: Изд-во РГЭУ (РИНХ), 2015	95
Л2.3	Герасимов А. Н., Громов Е. И., Скрипниченко Ю. С.	Эконометрика: учеб. пособие для студентов высш. учеб. заведений, обучающихся по напр. подгот. 38.03.01 "Экономика"	Ростов н/Д: Феникс, 2017	20
Л2.4		Хранилища данных. Лекция 1. Понятия о хранилищах. Презентация	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237105 неограниченный доступ для зарегистрированных пользователей
Л2.5		Хранилища данных. Лекция 3. Создание куба в SQL Server 2005. Презентация	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237113 неограниченный доступ для зарегистрированных пользователей
Л2.6		Хранилища данных. Лекция 4. Создание многомерного хранилища данных на основе MS SQL Server 2005. Презентация	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237114 неограниченный доступ для зарегистрированных пользователей
Л2.7		Хранилища данных. Лекция 6. Работа с OLAP срезами. Презентация	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237115 неограниченный доступ для зарегистрированных пользователей
Л2.8		Хранилища данных. Лекция 7. SQL Server – ProClarity. Презентация	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237116 неограниченный доступ для зарегистрированных пользователей
Л2.9		Хранилища данных	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237117 неограниченный доступ для зарегистрированных пользователей
Л2.10		Хранилища данных. Лекция 9. Обзор основных технологий и функциональных возможностей Crystal Analysis Professional 10.0. Презентация	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237118 неограниченный доступ для зарегистрированных пользователей
Л2.11	Рощина Я. М.	Основы моделирования экономического поведения домохозяйств на базе данных RLMS-HSE: лекции	Москва: Издательский дом Высшей школы экономики, 2015	http://biblioclub.ru/index.php?page=book&id=440284 неограниченный доступ для зарегистрированных пользователей
Л2.12		Прикладная эконометрика: журнал	Москва: Университет Синергия, 2018	https://biblioclub.ru/index.php?page=book&id=484968 неограниченный доступ для зарегистрированных пользователей

5.3 Профессиональные базы данных и информационные справочные системы
Для преподавания курса предполагается использование данных социологических опросов населения НАФИ (https://nafi.ru), ВЦИОМ (https://nafi.ru) и других, данные репрезентативных опросов населения России, таких как RLMS (https://www.hse.ru/rlms/), а также данные федеральной статистики ЕМИСС (www.fedstat.ru).
ИСС «КонсультантПлюс»
ИСС «Гарант» http://www.internet.garant.ru/
5.4. Перечень программного обеспечения
Libre Office, R, RStudio.
5.5. Учебно-методические материалы для студентов с ограниченными возможностями здоровья
При необходимости по заявлению обучающегося с ограниченными возможностями здоровья учебно-методические материалы предоставляются в формах, адаптированных к ограничениям здоровья и восприятия информации. Для лиц с нарушениями зрения: в форме аудиофайла; в печатной форме увеличенным шрифтом. Для лиц с нарушениями слуха: в форме электронного документа; в печатной форме. Для лиц с нарушениями опорно-двигательного аппарата: в форме электронного документа; в печатной форме.

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)
Помещения для всех видов работ, предусмотренных учебным планом, укомплектованы необходимой специализированной учебной мебелью и техническими средствами обучения:
- столы, стулья;
- персональный компьютер / ноутбук (переносной);
- проектор, экран / интерактивная доска.
Лабораторные занятия проводятся в компьютерных классах, рабочие места в которых оборудованы необходимыми лицензионными и/или свободно распространяемыми программными средствами и выходом в Интернет.

7. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)
Методические указания по освоению дисциплины представлены в Приложении 2 к рабочей программе дисциплины.

Приложение 1

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

1. Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания

1.1 Показатели и критерии оценивания компетенций:

ЗУН, составляющие компетенцию	Показатели оценивания	Критерии оценивания	Средства оценивания
ПК-4 способность формировать информационную базу бизнес-анализа, оценивать текущее состояние организации (объекта исследования) выявлять, и оценивать несоответствия между параметрами ее текущего и будущего состояний			
Знать базовые понятия и современные информационные технологии Big Data, а также языки визуального моделирования применяемые для целей бизнес-анализа (соответствует ПК-4.1).	Знает базовые понятия и технологии прогнозирования с использованием больших данных.	Верно использует модели и методы эконометрики для решения задачи на больших данных с учетом ограничений	С - собеседование, Т – тесты, ЗЗ – задание к зачету
Уметь анализировать большие массивы данных, характеризующие различные социально-экономические процессы, внутренние (внешние) факторы и условия, влияющие на деятельность организации; применять информационные технологии анализа Big Data в объеме, необходимом для целей бизнес-анализа (соответствует ПК-4.2).	Умеет определять массивы больших данных; анализировать кластеры больших данных. Строит различными способами прогнозы развития экономических процессов с использованием больших массивов данных.	Верно использует модели и методы эконометрики для решения задачи на больших данных с учетом ограничений	С - собеседование, З – задачи разного уровня, ЛР – лабораторная работа, ЗЗ – задание к зачету
Владеть современными	Владеет терминологией курса, методологией и методикой	Корректные методы для	З – задачи разного

технологиями создания и анализа больших данных, а также навыками применения информационных технологии в объеме, необходимом для целей бизнес-анализа (соответствует ПК-4.3).	прогнозирования.	решения задачи, адекватная модель, верная интерпретация результатов моделирования	уровня, ЛР – лабораторная работа, ЗЗ – задание к зачету

1.2. Шкала оценивания:

Текущий контроль успеваемости и промежуточная аттестация осуществляется в рамках накопительной балльно-рейтинговой системы в 100-балльной шкале.

Промежуточная аттестация осуществляется по следующей шкале:

- 50-100 баллов (зачтено)
- 0-49 баллов (не зачтено).

2. Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы

Задания к зачету

ЗАДАНИЕ К ЗАЧЕТУ №1

1. Технологии Business Intelligence и реляционные системы управления базами данных.
2. Интегрированные процессы. Идентификация авторегрессионной модели скользящего среднего.
3. Задача.

На встроенном датасете LifeCycleSavings предсказать значение sr на основе всех остальных переменных в этом датасете. Напишите команду, которая создаёт линейную регрессию с главными эффектами и всеми возможными взаимодействиями второго уровня. Сохраните модель в переменную model. Выпишите итоговую спецификацию модели и оцените ее значимость.

ЗАДАНИЕ К ЗАЧЕТУ № 2

1. Технологии обработки больших данных: NoSQL, MapReduce, Hadoop, R.
2. Коинтеграция. Анализ временных рядов.
3. Задача.

По 350 наблюдениям оценена модель зависимости заработной платы $wage_i$ (\$) от длительности обучения $schooling_i$ (годы) и опыта работы $experience_i$ (годы). Оцененная модель имеет вид: $\widehat{wage}_\iota = 400 + 25schooling_i + 60experience_i$. *ESS=130, TSS=210*. Исследователь решил добавить в модель образование родителей $mschooling_i$ и $fschooling_i$ (годы), после чего *ESS=180*. На уровне значимости 10% проверяя гипотезу о влиянии длительности обучения родителей на заработную плату их ребенка, укажите количество ограничений, которые приравнены к нулю в формулировке нулевой гипотезы? Чему равно значение наблюдаемой статистики?

ЗАДАНИЕ К ЗАЧЕТУ № 3

1. Понятие Большие данные. Роль цифровой информации в 21 веке.
2. Специальные методы анализа социально-политических и медиа процессов.
3. Задача.

По 400 наблюдениям оценена модель зависимости заработной платы $wage_i$ ($) от длительности обучения $schooling_i$ (годы) и опыта работы $experience_i$ (годы). Оцененная модель имеет вид:

$$\widehat{wage}_\iota = 400 + 25schooling_i + 60experience_i.$$ ESS=125, TSS=200.

Исследователь решил добавить в модель образование родителей $mschooling_i$ и $fschooling_i$ (годы), после чего ESS=175. На уровне значимости 1% проверяя гипотезу о влиянии длительности обучения родителей на заработную плату их ребенка, определите чему равно наблюдаемое значение тестовой статистики?

ЗАДАНИЕ К ЗАЧЕТУ № 4

1. Базовые принципы обработки больших данных.
2. Статистические оценки параметров. Доверительные области.
3. Задача.

Исследуется зависимость среднедушевого потребления алкоголя по странам мира от различных факторов.

Модель 1:

ALCOi=β1+β2GDPi+β3MUSLi+β4BUDDi+β5HINDUi+εi,

где ALCOi — среднедушевое потребление чистого спирта на человека (л), GDPi — ВВП на душу населения (долларов США), MUSLi, BUDDi, HINDUi — доли населения исповедующего, соответственно, мусульманство, буддизм и индуизм (в % от общей численности населения). В ходе МНК-оценивания модели на основе данных о 50 странах получены следующие результаты: сумма квадратов остатков ESS=200, объясненная сумма квадратов RSS=300.

Также для проверки гипотезы о том, что религия не оказывает существенного влияния на потребление алкоголя, были оценены параметры второй модели:

Модель №2:

ALCOi=β1+β2GDPi+εi.

Во второй модели, по сравнению с первой, значение RSS увеличилось на 100. Сколько составит скорректированный R^2 во второй модели?

ЗАДАНИЕ К ЗАЧЕТУ № 5

1. Предварительный анализ данных.
2. Основные возможности анализа больших данных в R.
3. Задача.

Вопросы этого задания основаны на следующем эксперименте: 400 водителей, выбранных случайным образом, попросили пройти специальный тест на вождение автомобилем. Для каждого водителя были собраны следующие данные: Pass — фиктивная переменная, равная единице, если водитель сдал тест, Male — фиктивная переменная, равная единице, если водитель мужчина, и равная 0, если водитель женщина, Experience — опыт вождения автомобилем (в годах). В таблице представлены результаты семи моделей, оцененных на основе имеющихся данных.

Dependent Variable: Pass							
	Probit	Logit	Linear Probability	Probit	Logit	Linear Probability	Probit
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Experience	0.031 (0.009)	0.040 (0.016)	0.006 (0.002)				0.041 (0.156)
Male				–0.333 (0.161)	–0.622 (0.303)	–0.071 (0.034)	–0.174 (0.259)
Male*Experience							–0.015 (0.019)
Constant	0.712 (0.126)	1.059 (0.221)	0.774 (0.034)	1.282 (0.124)	2.197 (0.242)	0.900 (0.022)	0.806 (0.200)

Используйте результаты из колонки (1). Каков предельный эффект дополнительного года опыта для Джейн – женщины с 10-летним опытом вождения?

ЗАДАНИЕ К ЗАЧЕТУ № 6

1. Неметрические методы. Кластерный анализ. Дискриминантный анализ.
2. Коинтеграция. Анализ временных рядов.
3. Задача.

   Используем встроенный набор данных mtcars в RStudio. Сохраните в переменную логистическую регрессионную модель, где в качестве зависимой переменной выступает тип коробки передач (am), в качестве предикторов переменные disp, vs, mpg. Значения коэффициентов регрессии сохраните в переменную log_coef. Выпишите полученную спецификацию модели. Оцените ее качество.

ЗАДАНИЕ К ЗАЧЕТУ № 7

1. Основные возможности анализа больших данных в R.
2. Специальные методы анализа социально-политических и медиа процессов.
3. Задача.

   На встроенных в R данных npk, иллюстрирующими влияние применения различных удобрений на урожайность гороха (yield). Нашей задачей будет выяснить, существенно ли одновременное применение азота (фактор N) и фосфата (фактор P). Примените дисперсионный анализ, где будет проверяться влияние фактора применения азота (N), влияние фактора применения фосфата (P) и их взаимодействие. В ответе укажите p-value для взаимодействия факторов N и P.

ЗАДАНИЕ К ЗАЧЕТУ № 8

1. Теория моментов.
2. Многомерное шкалирование. Классическая модель многомерного шкалирования.
3. Задача.

   В переменной df сохранен subset данных mtcars только с переменными "wt", "mpg", "disp", "drat", "hp". Воспользуйтесь множественным регрессионным анализом, чтобы предсказать вес машины (переменная "wt"). Выберите такую комбинацию независимых переменных (из "mpg", "disp", "drat", "hp"), чтобы значение R^2 adjusted было наибольшим. Взаимодействия факторов учитывать не надо.

ЗАДАНИЕ К ЗАЧЕТУ № 9

1. Специальные методы анализа социально-политических и медиа процессов.
2. Дисперсионный анализ влияния качественных факторов. Ранговые методы.
3. Задача.

   Воспользуйтесь встроенным датасетом attitude, чтобы предсказать рейтинг (rating) по переменным complaints и critical. Каково t-значение для взаимодействия двух факторов? Разделителем целой и дробной части в ответе должна быть запятая!

ЗАДАНИЕ К ЗАЧЕТУ № 10

1. Прогнозирование с моделями временных рядов. Доверительные интервалы прогноза.
2. Технологии обработки больших данных: NoSQL, MapReduce, Hadoop, R.
3. Задача.

   На встроенном датасете LifeCycleSavings предсказать значение sr на основе всех остальных переменных в этом датасете. Напишите команду, которая создаёт линейную регрессию с главными эффектами и всеми возможными взаимодействиями второго уровня. Сохраните модель в переменную model. Выпишите итоговую спецификацию модели и оцените ее качество.

Критерии оценивания:

Максимальное количество баллов – 100.

Задание к зачету содержит 2 вопроса и 1 задачу, баллы и критерии оценивания по которым приведены выше. Баллы выставляются по каждому заданию в отдельности и суммируются.

Каждый теоретический вопрос оценивается отдельно, максимально в 24 балла.

Критерии оценивания отдельного вопроса:

- 13-24 балла. Ответ на вопрос верный; продемонстрировано наличие глубоких исчерпывающих / твердых и достаточно полных знаний, грамотное и логически стройное изложение материала при ответе, возможны отдельные погрешности и ошибки, уверенно исправленные и после дополнительных вопросов; продемонстрировано наличие глубоких исчерпывающих / твердых и достаточно полных знаний, грамотное и логически стройное изложение материала при ответе.
- 0-12 баллов. Ответ на вопрос лишь частично верен, продемонстрирована неточность и неуверенность ответов на дополнительные и наводящие вопросы, либо ответ на вопрос не верен, продемонстрирована неуверенность и неточность ответов на дополнительные и наводящие вопросы.

Задача оценивается максимально в 52 балла:

Критерии оценивания задачи:

- 27-52 балла. Задача решена в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов; либо задача решена в полном объеме с небольшими погрешностями, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.
- 0-26 баллов. Задача решена частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с погрешностями либо задача не решена или решена частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

Зачет выставляется на основании итоговой суммы баллов, набранных студентом:

- 50-100 баллов «зачтено»;
- 0-49 баллов «не зачтено».

Тесты письменные и/или компьютерные
(Тестовые задания к экзамену и текущему контролю знаний)

Банк тестов

1. Банк тестов по модулям.

Модуль 1 «Сбор, хранение и анализ больших данных»

1.1 Задача классификации сводится к …

а) нахождения частых зависимостей между объектами или событиями;

б) определения класса объекта по его характеристикам;

в) определение по известным характеристикам объекта значение некоторого его параметра;

г) поиска независимых групп и их характеристик во всем множестве анализируемых данных.

1.2. Целью поиска ассоциативных правил является …

а) нахождения частых зависимостей между объектами или событиями;

б) определения класса объекта по его характеристикам;

в) определение по известным характеристикам объекта значение некоторого его параметра;

г) поиска независимых групп и их характеристик во всем множестве анализируемых данных.

1.3. Очистка данных — …

а) комплекс методов и процедур, направленных на устранение причин, мешающих корректной обработке: аномалий, пропусков, дубликатов, противоречий, шумов и т.д.;

б) процесс дополнения данных некоторой информацией, позволяющей повысить эффективность развязку аналитических задач;

в) объект, содержащий структурированные данные, которые могут оказаться полезными для решения аналитической задачи;

г) комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразования в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.

1.4 Обогащение — …

а) комплекс методов и процедур, направленных на устранение причин, мешающих корректной обработке: аномалий, пропусков, дубликатов, противоречий, шумов и т.д. ;
б) процесс дополнения данных некоторой информацией, позволяющей повысить эффективность развязку аналитических задач;
в) объект, содержащий структурированные данные, которые могут оказаться полезными для решения аналитической задачи;
г) комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразования в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.
1.5. Транзакция — …
а) некоторый набор операций над базой данных, который рассматривается как единственное завершенное, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных;
б) разновидность систем хранения, ориентирована на поддержку процесса анализа данных целостность, обеспечивает, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов;
в) высокоуровневые средства отражения информационной модели и описания структуры данных;
г) это установление зависимости дискретной выходной переменной от входных переменных.
1.6. Аналитическая платформа — …
а) специализированное программное решение (или набор решений), которое включает в себя все инструменты для извлечения закономерностей из сырых данных;
б) это группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов;
в) набор данных, каждая запись которого представляет собой учебный пример, содержащий заданное входное влияние, что и отвечает ему правильный выходной результат;
г) подразделение искусственного интеллекта, изучающее методы построения алгоритмов, способных обучаться на данных.
1.7. Консолидация — …
а) комплекс методов и процедур, направленных на устранение причин, мешающих корректной обработке: аномалий, пропусков, дубликатов, противоречий, шумов и т.д.;
б) процесс дополнения данных некоторой информацией, позволяющей повысить эффективность развязку аналитических задач;
в) объект, содержащий структурированные данные, которые могут оказаться бесполезными для решения аналитической задачи;
г) комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразования в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.
1.8. Виды физической неопределенности данных:
а) неточность измерений значений определенной величины, выполняемых физическими приборами; случайность (или наличие во внешней среде нескольких возможностей, каждая из которых случайным образом может стать действительностью);
б) неопределенность значений слов (многозначность, размытость, непонятность, нечеткость); неоднозначность смысла фраз (синтаксическая и семантическая);
в) случайность (или наличие во внешней среде нескольких возможностей, каждая из которых случайным образом может стать действительностью);
г) неопределенность значений слов (многозначность, размытость, неясность, нечеткость);
д) неоднозначность смысла фраз (синтаксическая и семантическая).
1.9. Метаданные — …
а) некоторый набор операций над базой данных, который рассматривается как единственное завершенное, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных;
б) разновидность систем хранения, ориентирована на поддержку процесса анализа данных целостность, обеспечивает, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов;
в) высокоуровневые средства отражения информационной модели и описания структуры данных;
г) это установление зависимости дискретной выходной переменной от входных переменных.
1.10. Классификация — …

а) некоторый набор операций над базой данных, который рассматривается как единственное завершенное, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных;
б) разновидность систем хранения, ориентированная на поддержку процесса анализа данных, обеспечивает целостность, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов;
в) высокоуровневые средства отражения информационной модели и описания структуры данных;
г) это установление зависимости дискретной выходной переменной от входных переменных.
1.11. Регрессия — …
а) это установление зависимости непрерывной выходной переменной от входных переменных;
б) эта группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов;
в) выявление закономерностей между связанными событиями;
г) это установление зависимости дискретной выходной переменной от входных переменных.
1.12. Кластеризация — …
а) это установление зависимости непрерывной выходной переменной от входных переменных;
б) эта группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов;
в) выявление закономерностей между связанными событиями;
г) это установление зависимости дискретной выходной переменной от входных переменных.
1.13. Ассоциация — …
а) это установление зависимости непрерывной выходной переменной от входных переменных;
б) эта группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов;
в) выявление закономерностей между связанными событиями;
г) это установление зависимости дискретной выходной переменной от входных переменных.
1.14. Машинное обучение — …
а) специализированное программное решение (или набор решений), которое включает в себя все инструменты для извлечения закономерностей из сырых данных;
б) это группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов;
в) набор данных, каждая запись которого представляет собой учебный пример, содержащий заданное входное влияние, и соответствующий ему правильный выходной результат;
г) подразделение искусственного интеллекта, изучающее методы построения алгоритмов, способных обучаться на данных.
1.15 Обучающая выборка — …
а) это группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов;
б) набор данных, каждая запись которого представляет собой учебный пример, содержащий заданное входное влияние, и соответствующий ему правильный выходной результат;
в) набор данных, каждая запись которого представляет собой учебный пример, содержащий заданный входной влияние, что и обеспечивает ему правильный выходной результат;
г) выявление в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.
1.16. Укажите фактор, способствовавший появлению тренда больших данных
а) маркетинговые кампании крупных корпораций;
б) снижение издержек на хранение данных;
в) появление новых технологий обработки потоковых данных;
г) выпуск баз данных с обработкой данных в памяти.
1.17. Отметьте значимые события, повлиявшие на формирование тренда больших данных:
а) разработка Hadoop;
б) изобретение принципа MapReduce;
в) разработка языка Pyhton;
г) победа Deepblue в матче с Г.Каспаровым.
1.18. Выберите верный ответ:
а) большие данные – это обработка или хранение более 1 Тб информации;
б) проблема больших данных – это такая проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна;
в) большие данные – это огромная PR-акция крупных вендоров и не более того;

г) большие данные – это явление, когда цифровые данные наиболее полно представляют изучаемый объект.
1.19. Выберите неверный ответ:
а) большие данные – это данные объёма свыше 1 Тб;
б) проблема больших данных – это проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна;
в) большие данные – это тренд в области ИТ, подогреваемый маркетинговыми кампаниями крупных вендоров;
г) большие данные как правило не структурированы.
1.20. Отметьте те из вариантов, в которых данные структурированы:
а) данные о продажах компании, представленные в виде помесячных отчётов в формате MS Word;
б) таблица с ежедневными показаниями температуры помещения за год в файле формата csv;
в) текст педагогической поэмы А.С. Макаренко, представленный в формате PDF;
г) библиотека фильмов, представленных в формате mpeg4 на одном жестком диске.
1.21. Перечислите четыре основных характеристики Big Data:
а) Virtualization, Volume, Variability, Vehicle;
б) Variety, Velocity, Volume, Value;
в) Verification, Volume, Velocity, Visualization;
г) Video, Value, Variety, Volume.
1.22. Выберите неверное высказывание:
а) большие объёмы данных приводят к слабой их структуризации, поэтому появляется такое разнообразие данных;
б) увеличившаяся производительность телекоммуникационных каналов привела к росту объёмов передаваемой информации;
в) удешевление систем хранения на единицу информации привело к росту рынка больших данных;
г) большое разнообразие источников данных
1.23. Отметьте неверное понимание Variety в контексте характеристик Big Data:
а) высокая скорость генерирования данных;
б) разные типы данных в колонках таблиц реляционных СУБД;
в) разнообразие отраслей, являющихся источниками данных;
г) разнообразие типов данных, включающих в себя структурированные, полуструктурированные и неструктурированные.
1.24. Принцип MapReduce состоит в том, чтобы:
а) производить вычисления на узлах, где информация изначально была сохранена;
б) использовать вычислительные мощности систем хранения;
в) использовать функциональное программирование для решения задач массивно-параллельной обработки.
1.25. Выберите одно неверное высказывание про MapReduce:
а) интерфейс для массово-параллельной обработки данных, где вычисления производятся на узлах, где информация изначально была сохранена;
б) MapReduce – это две операции: распределения и сборки данных;
в) MapReduce был придуман разработчиками Hadoop;
г) MapReduce был анонсирован разработчиками Google.
1.26. Какие из следующих технологий СУБД не используют принцип MapReduce:
а) Hadoop;
б) Cassandra;
в) HDInsight;
г) Redis.
1.27. Какие СУБД полностью полагаются на оперативную память при хранении информации:
а) Oracle Exalytics;
б) SAP HANA;
в) BigTable;
г) HBase.
1.28. В чём преимущество колоночно-ориентированных СУБД?
а) они позволяют выполнять более сложные SQL-запросы по сравнению с реляционными СУБД;
б) они позволяют динамически дополнять содержание записей новыми полями;
в) они имеют более гибкие возможности аналитики;

г) они позволяют эффективно делать межколоночные сравнения.
1.29. На каком из этапов процесса CRISP-DM происходит проверка гипотез?
а) понимание бизнеса (Business understanding);
б) понимание данных (Data Understanding);
в) моделирование (Modeling);
г) оценка (Evaluation).
1.30. Вы являетесь владельцем и аналитиком в компании из 10 человек, в которой требуется проанализировать продажи за 1 год (1 млн. продаж). Какие из этапов CRISP-DM можно опустить:
а) понимание бизнеса (Business understanding);
б) подготовка данных (Data Preparation);
в) моделирование (Modeling);
г) оценка (Evaluation).
1.31. Пример благоразумного использования Hadoop:
а) анализ 10 Гб данных;
б) ежедневное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт.);
в) посекундное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт.);
г) построение графика пульса пациента в реальном времени.
1.32. Начиная с каких размеров данных обоснованно применение кластера Hadoop для хранения данных?

_____

1.33. Hadoop – это:
а) набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах;
б) распределённая СУБД, позволяющая обрабатывать большие данные;
в) язык выполнения заданий в парадигме MapReduce;
г) распределённая файловая система, предназначенная для хранения файлов большого объёма.
1.34. Подразделение искусственного интеллекта, изучающее методы построения алгоритмов, способных обучаться на данных – это…

Модуль 2 «Методы и модели анализа больших данных»

2.1. Модели классификации описывают …
а) правила или набор правил, в соответствии с которыми можно отнести описание любого нового объекта к одному из классов;
б) функции, которые позволяют прогнозировать изменения непрерывных числовых параметров;
в) функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме;
г) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.
2.2. Модели исключений описывают …
а) исключительные ситуации в записях, которые резко отличаются по произвольному признаку от основного множества записей;
б) ограничения на данные анализируемого массива;
в) закономерности между связанными событиями;
г) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.
2.3. Итоговые модели обнаружат …
а) исключительные ситуации в записях, которые резко отличаются по произвольному признаку от основного множества записей;
б) ограничения на данные анализируемого массива;
в) закономерности между связанными событиями;
г) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.
2.4. Задача регрессии сводится к …
а) нахождению частных зависимостей между объектами или событиями;
б) определения класса объекта по его характеристикам;
в) определение по известным характеристикам объекта значение некоторого его параметра;
г) поиска независимых групп и их характеристик во всем множестве анализируемых данных.
2.5. Модели последовательностей описывают …
а) правила или набор правил, в соответствии с которыми можно отнести описание любого нового объекта к одному из классов;

б) функции, которые позволяют прогнозировать изменения непрерывных числовых параметров;
в) функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме;
г) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.
2.6. Модели ассоциации проявляют …
а) исключительные ситуации в записях, которые резко отличаются по произвольному признаку от основной множества записей;
б) ограничения на данные анализируемого массива;
в) закономерности между связанными событиями;
г) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.
2.7. Регрессионные модели описывают …
а) правила или набор правил в соответствии с которыми можно отнести описание любого нового объекта к одному из классов;
б) функции, которые позволяют прогнозировать изменения непрерывных числовых параметров;
в) функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме;
г) группы, на которые можно разделить объекты, данные о которых подвергаются анализа.
2.8. Задача кластеризации заключается в …
а) нахождении частых зависимостей между объектами или событиями;
б) определении класса объекта по его характеристикам;
в) определении по известным характеристикам объекта значение некоторого его параметра;
г) поиске независимых групп и их характеристик во всем множестве анализируемых данных.
2.9. Ошибка обучения — …
а) это ошибка, допущенная моделью на учебном множестве;
б) это ошибка, полученная на тестовых примерах, то есть, что вычисляется по тем же формулам, но для тестового множества;
в) имена, типы, метки и назначения полей исходной выборки данных;
г) набор данных, каждая запись которого представляет собой учебный пример, содержащий заданное входное влияние, и соответствующий ему правильный выходной результат.
2.10. Ошибка обобщения — …
а) это ошибка, допущенная моделью на учебном множестве;
б) это ошибка, полученная на тестовых примерах, то есть, вычисляется по тем же формулам, но для тестового множества;
в) имена, типы, метки и назначения полей исходной выборки данных;
г) набор данных, каждая запись которого представляет собой учебный пример, содержащий заданное входное влияние, и соответствующий ему правильный выходной результат.
2.11. Аналитик это …
а) специалист в области анализа и моделирования;
б) специалист в предметной области;
в) человек, решающий определенные задачи;
г) человек, который имеет опыт в программировании.
2.12. Задача классификации сводится к …
а) нахождению частых зависимостей между объектами или событиями;
б) определению класса объекта по его характеристикам;
в) определению по известным характеристикам объекта значение некоторого его параметра;
г) поиску независимых групп и их характеристик в всем множестве анализируемых данных.
2.13. «Песочница» в аналитическом процессе:
а) для чего аналитику необходима «песочница»?
б) для высокопроизводительной аналитики за счёт использования оперативной памяти и inDB операций;
в) для хранения всех полученных от заказчика данных;
г) для построения отчётов о результатах анализа;
д) для снижения затрат, связанных с репликацией данных.
2.14. Какие из следующих средств разумно использовать для анализа данных, представленных единственным csv-файлом размера более 100Гб:
а) Hadoop;
б) Data Warehouse;
в) «Песочница»;

г) Python.
2.15. Выберите верное утверждение:
а) Data Warehouse создаются для проверки гипотез при анализе больших данных;
б) «Песочница» используется для снижения нагрузки на основной Data Warehouse;
в) каждый Data Warehouse должен содержать «песочницу»;
г) «Песочница» необходима для любого процесса аналитики.
2.16. Ошибка, полученная на тестовых примерах (вычисляется по тем же формулам, но для тестового множества) – это ошибка…
2.17. Функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме описывают модели…
2.18. Закономерности между связанными событиями описывают модели…

Критерии оценивания:
Максимальный балл – 20.
Число вопросов - 20. Ответ на каждый вопрос оценивается максимум в 1 балл.
Критерии оценивания 1 вопроса:
0,84-1,0 балла выставляется студенту, если изложенный материал фактически верен, продемонстрированы глубокие исчерпывающие знания в объеме пройденной программы в соответствии с поставленными программой курса целями и задачами обучения, изложение материала при ответе грамотное и логически стройное;
0,67-0,83 балла выставляется студенту, если продемонстрированы твердые и достаточно полные знания в объеме пройденной программы дисциплины в соответствии с целями обучения; материал изложен достаточно полно с отдельными логическими и стилистическими погрешностями;
0,5-0,66 балла выставляется студенту, если продемонстрированы твердые знания в объеме пройденного курса в соответствие с целями обучения, ответ содержит отдельные ошибки, уверенно исправленные после дополнительных вопросов;
0-0,49 балла выставляется студенту, если ответ не связан с вопросом, допущены грубые ошибки в ответе, продемонстрированы непонимание сущности излагаемого вопроса, неуверенность и неточность ответов на дополнительные и наводящие вопросы.

Вопросы для собеседования

Модуль 1 «Сбор, хранение и анализ больших данных»

1. Что такое Big Data? Понятие, сущность и ключевые признаки больших данных.
2. Роль и место больших данных в решении аналитических и исследовательских задач профессиональной деятельности.
3. Методы и средства анализа Big Data.
4. Используемые программы для анализа Big Data.
5. Технологии хранения больших данных.
6. Определение источника больших данных.
7. Исследование источника данных.
8. Что такое хранилище данных?
9. Процесс анализа больших данных.
10. Технологии анализа больших данных.
11. Научные проблемы в области больших данных.
12. Технологии и инструменты больших данных.
13. Apache Hadoop. Storm – система потоковой обработки.
14. Язык программирования R.
15. Аналитика больших данных как корпоративный проект.
16. Сущность и принцип работы аналитической платформы Deductor Academic.
17. Основные функции и инструменты аналитической платформы Deductor Academic для целей анализа и исследования социально-экономических процессов и явлений в деятель-ности предприятий.

18. Моделирование социально-экономических процессов и явлений в деятельности пред-приятий с помощью платформы Deductor Academic.
19. Инструментарий прикладного компьютерного анализа и моделирования в Deductor Academic.

Модуль 2 «Методы и модели анализа больших данных»

20. Прогнозирование и предвидение в социально-политических и медиа процессах.
21. Методы прогнозирования, использующие большие данные.
22. Что такое OLAP-система?
23. Чем OLAP-системы отличаются от Big Data?
24. Техники больших данных.
25. Консолидация данных.
26. Визуализация.
27. Классификация.
28. Кластеризация.
29. Регрессионный анализ.
30. Анализ ассотиативных правил.
31. Нейронные сети. Экспоненциальное сглаживание Холта с корректировкой тренда в моделях временных рядов.
32. Интеллектуальный анализ данных (Data Mining).
33. Задачи и методы интеллектуального анализа больших данных.
34. Инструменты Data Mining.

Общее число вопросов на собеседовании 3. Каждый вопрос оценивается отдельно, максимально в 5 балла.

Критерии оценивания отдельного вопроса:

- 2,5-5,4 балла. Ответ на вопрос верный; продемонстрировано наличие глубоких исчерпывающих / твердых и достаточно полных знаний, грамотное и логически стройное изложение материала при ответе, возможны отдельные погрешности и ошибки, уверенно исправленные и после дополнительных вопросов; продемонстрировано наличие глубоких исчерпывающих / твердых и достаточно полных знаний, грамотное и логически стройное изложение материала при ответе.
- 0-2,4 балла. Ответ на вопрос лишь частично верен, продемонстрирована неточность и неуверенность ответов на дополнительные и наводящие вопросы, либо ответ на вопрос не верен, продемонстрирована неуверенность и неточность ответов на дополнительные и наводящие вопросы.

Комплект разноуровневых задач (заданий)

1 Задачи репродуктивного уровня

Задача 1. Исследователь анализирует зависимость потребления (c) от располагаемого дохода (y) на основе простой эмпирической модели: $c_i=\beta y_i+\varepsilon_i$, ε_i - независимые нормально распределенные случайные величины с нулевым математическим ожиданием и дисперсией $V(\varepsilon_i) = a^2 \cdot y_i^2$.

Исследователь собрал данные о двух тысячах домашних хозяйств и осуществил следующие предварительные расчёты:

$\sum_{i=1}^{2000} y_i = 2000;\ \sum_{i=1}^{2000} c_i = 1000;\ \sum_{i=1}^{2000} y_i^2 = 1450;\ \sum_{i=1}^{2000} y_i c_i = 950;\ \sum_{i=1}^{2000} \frac{y_i}{c_i} = 1050;\ \sum_{i=1}^{2000} \frac{c_i}{y_i} = 1550.$

Используя те из доступных данных, которые вам необходимы, вычислите эффективную оценку предельной склонности к потреблению.

Задача 2. Исследуется зависимость среднедушевого потребления алкоголя по странам мира от различных факторов.

Модель 1:

$$ALCO_i=\beta_1+\beta_2 GDP_i+\beta_3 MUSL_i+\beta_4 BUDD_i+\beta_5 HINDU_i+\varepsilon_i,$$

где $ALCO_i$ — среднедушевое потребление чистого спирта на человека (л), GDP_i — ВВП на душу населения (долларов США), $MUSL_i$, $BUDD_i$, $HINDU_i$ — доли населения исповедующего, соответственно,

мусульманство, буддизм и индуизм (в % от общей численности населения). В ходе МНК-оценивания модели на основе данных о 50 странах получены следующие результаты: сумма квадратов остатков *ESS=200*, объясненная сумма квадратов *RSS=300*.

Также для проверки гипотезы о том, что религия не оказывает существенного влияния на потребление алкоголя, были оценены параметры второй модели:

Модель №2:

$$ALCO_i=\beta_1+\beta_2 GDP_i+\varepsilon_i.$$

Во второй модели, по сравнению с первой, значение *RSS* изменилось на100. Сколько составит корректированный R^2 во второй модели?

Задача 3.

По 100 наблюдениям Вениамин оценил зависимость количества решённых им за вечер задач по эконометрике (problems) от количества съеденных булочек с яблоками (applepie) и мясом (meatpie):

$$\widehat{problems}_i = 2 + 0.5 \cdot applepie_i + 0.7 \cdot meatpie_i$$

Стандартные ошибки коэффициентов при (applepie) и (meatpie) равны, соответственно, 0.1 и 0.5.
Для проверки гипотезы о том, что эффект от булочек с яблоками и булочек с мясом одинаковый Вениамину стоит

Задача 4.

По 30 наблюдениям было оценено следующее уравнение регрессии (в скобках указаны стандартные отклонения оценок коэффициентов):

$$\widehat{y_i} = \underset{(1,0)}{1,5} - \underset{(0,4)}{0,9} \cdot x_i^{(1)} + \underset{(0,01)}{0,04} \cdot x_i^{(2)} + \underset{(0,02)}{0,09} \cdot x_i^{(3)} + \underset{(0,6)}{2,0} \cdot x_i^{(4)},\ R^2 = 0,59$$

Проверьте (при уровне значимости 5%) гипотезу о том, что все коэффициенты при переменных уравнения одновременно равны нулю.

Задача 5.

По 2040 наблюдениям Вениамин оценил модель зависимости стоимости цены квартиры $price_i$ (в 1000\$) от расстояния до ближайшего метро $metrdist_i$:
$\widehat{price}_i = 144.2848 - 2.0682 \cdot metrdist_i$. При построении 95% доверительного интервала для $\mathrm{E}(price_f|metrdist = 15)$ Вениамин столкнулся с проблемой рассчёта.

Чему равна $\widehat{\mathrm{Var}}(\widehat{price}_f|X)$, если $\hat{\sigma}^2 = 2630.364$, а ковариационная матрица имеет следующий вид: $\widehat{\mathrm{Var}}(\hat{\beta}|X) = \begin{pmatrix} & (Intercept) & metrdist \\ (Intercept) & 7.128 & -0.719 \\ metrdist & -0.719 & 0.0886 \end{pmatrix}$.

Округляйте до одного знака после запятой.

Задача 6.

По 2040 наблюдениям Вениамин оценил модель зависимости стоимости цены квартиры $price_i$ (в 1000\$) от расстояния до ближайшего метро $metrdist_i$:
$\widehat{price}_i = 144.2848 - 2.0682 \cdot metrdist_i$. При построении 90% предиктивного интервала для $price_f\ (metrdist = 15)$ он столкнулся с проблемой расчёта.

Чему равна $\widehat{\mathrm{Var}}(price_f - \widehat{price}_f|X)$, если $\hat{\sigma}^2 = 2630.364$, а ковариационная матрица имеет следующий вид: $\widehat{\mathrm{Var}}(\hat{\beta}|X) = \begin{pmatrix} & (Intercept) & metrdist \\ (Intercept) & 7.128 & -0.719 \\ metrdist & -0.719 & 0.0886 \end{pmatrix}$.

Округляйте до одного знака после запятой.

2 Задачи реконструктивного уровня

Задача 1. Вопросы этого задания основаны на следующем эксперименте: 400 водителей, выбранных случайным образом, попросили пройти специальный тест на вождение автомобилем. Для каждого водителя были собраны следующие данные: Pass — фиктивная переменная, равная единице, если водитель сдал тест, Male — фиктивная переменная, равная единице, если водитель мужчина, и равная 0, если водитель женщина, Experience — опыт вождения автомобилем (в

годах). В таблице представлены результаты семи моделей, оцененных на основе имеющихся данных.

Dependent Variable: Pass							
	Probit (1)	Logit (2)	Linear Probability (3)	Probit (4)	Logit (5)	Linear Probability (6)	Probit (7)
Experience	0.031 (0.009)	0.040 (0.016)	0.006 (0.002)				0.041 (0.156)
Male				–0.333 (0.161)	–0.622 (0.303)	–0.071 (0.034)	–0.174 (0.259)
Male*Experience							–0.015 (0.019)
Constant	0.712 (0.126)	1.059 (0.221)	0.774 (0.034)	1.282 (0.124)	2.197 (0.242)	0.900 (0.022)	0.806 (0.200)

Используйте результаты из колонки (1). Каков предельный эффект дополнительного года опыта для Джейн – женщины с 10-летним опытом вождения?

Задача 2. На встроенном датасете LifeCycleSavings предсказать значение sr на основе всех остальных переменных в этом датасете. Напишите команду, которая создаёт линейную регрессию с главными эффектами и всеми возможными взаимодействиями второго уровня. Сохраните модель в переменную model. Выпишите итоговую спецификацию модели и оцените ее значимость.

Задача 3. Вениамин исследует влияние различных факторов на заработную плату людей, проживающих в России. По 1,000 наблюдений он оценил регрессию заработной платы ($wage_i$, в тысячах рублей) от образования человека ($educ_i$, в годах), образования матери этого человека ($mother.educ_i$, в годах) и дамми на опыт работы ($exper_i$: 1 - если более 5 лет, 0 - если менее 5 лет). Вениамин получил следующие результаты:

$$\widehat{wage}_i = 10 + 20 \cdot educ_i + 5 \cdot mother.educ_i + 15 \cdot exper_i$$

У Вениамина есть опасения, что в данных есть мультиколлинеарность. Чтобы проверить свою догадку, он оценил вспомогательные регрессии и получил:

$$\widehat{educ}_i = 11 + 2 \cdot mother.educ_i - 5 \cdot exper_i, R^2 = 0.97$$

$$\widehat{mother.educ}_i = 10 + 10 \cdot educ_i + 3 \cdot exper_i, R^2 = 0.93$$

$$\widehat{exper}_i = -8 \cdot educ_i + 9 \cdot mother.educ_i, R^2 = 0.3$$

Чему равен коэффициент вздутия дисперсии в вспомогательной регрессии образования матери на остальные объясняющие переменные? Ответ округляйте до двух знаков после запятой.

Задача 4. Исследователь анализирует влияние вступления в брак на уровень доходов. Он собрал данные за четыре года о тысяче работников обоих полов, часть из которых в течение рассматриваемого периода вступили в брак. В качестве объясняющей исследователь использовал фиктивную переменную *Одинокий*, которая равна единице для тех работников, которые в данном году не женаты (не замужем). В качестве контрольных переменных он использовал переменные *Возраст* (возраст в годах) и *Образование* (число лет обучения в годах). Исследователь оценил четыре уравнения (см. таблицу): первые два оценены обычным МНК, третье и четвертое оценены при помощи модели с фиксированными эффектами (внутригрупповое преобразование).

Таблица 1. Результаты оценки моделей. Зависимая переменная — логарифм заработной платы работника.

Модель	**Модель 1**	**Модель 2**	**Модель 3**	**Модель 4**
Метод оценивания	МНК	МНК	FE	FE
Возраст	0,80 (0,10)	0,78 (0,09)	0,67 (0,11)	0,68 (0,11)
Число лет обучения	1,20 (0,24)	1,41 (0,23)	0,91 (0,50)	0,90 (0,52)
Одинокий	–0,05 (0,02)	–0,04 (0,02)	–0,02 (0,01)	–0,03 (0,01)

Индивидуальные эффекты	Нет	Нет	Да	Да
Фиктивные переменные времени	Нет	Да	Нет	Да
Число наблюдений	4000	4000	4000	4000
R-квадрат	0,612	0,723	0,520	0,521
P-значение теста на отсутствие индивидуальных эффектов	—	—	0,001	0,002
P-значение теста на равенство нулю коэффициентов при фиктивных переменных времени	—	0,090	—	0,170

В скобках под оценками коэффициентов указаны робастные стандартные ошибки. В случае МНК представлен скорректированный R-квадрат, в случае модели с фиксированными эффектами within-R-квадрат.

Какую из четырех моделей следует выбрать в соответствии с доступной информацией?

3 Задачи творческого уровня

Задача 1. Используя открытые статистические данные, постройте регрессионную модель на больших данных. Приведите теоретическое обоснование вида и структуры модели. Оцените ее качество. Сделайте выводы.

Задача 2. Проведите исследование, чтобы выявить факторы, влияющие на доход людей, на реальных данных Российского мониторинга экономического положения и здоровья населения (РМЭЗ, он же RLMS). Воспользуйтесь для этого репрезентативной выборкой по индивидам за 2014 год - волна 23 (нужный файл называется r23i_os26a.sav). Для загрузки данных в R воспользуйтесь пакетом rlms.

Набор включает по одной задаче каждого уровня (суммарно максимально 10 баллов).
Задача репродуктивного уровня оценивается максимально в 2 балла:
Критерии оценивания задачи:

- 1-2 балла. Задача решена в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов; либо задача решена в полном объеме с небольшими погрешностями, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.
- 0-0,9 балла. Задача решена частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с погрешностями либо задача не решена или решена частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

Задача реконструктивного уровня оценивается максимально в 3 балла:
Критерии оценивания задачи:

- 1,5-3 балла. Задача решена в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов; либо задача решена в полном объеме с небольшими погрешностями, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.
- 0-1,4 балла. Задача решена частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с погрешностями либо задача не решена или решена частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

Задача творческого уровня оценивается максимально в 5 баллов:
Критерии оценивания задачи:

- 2,5-5 баллов. Задача решена в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов; либо задача решена в полном объеме с небольшими погрешностями, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.
- 0-2,4 баллов. Задача решена частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с погрешностями либо задача не решена или решена частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

Лабораторные работы

1.Тематика лабораторных работ по разделам и темам

Модуль 1 «Сбор, хранение и анализ больших данных»

1. Введение в R. Работа с электронными таблицами больших данных.
2. RStudio: визуализация Big Data, фиктивные переменные, прогнозы, проверка гипотез и ловушка дамми-переменных.
3. Большие данные и скрипты. Проблемы регрессионного анализа больших данных (мультиколлинеарность и гетероскедастичность) в RStudio.

Модуль 2 «Методы и модели анализа больших данных»

4. Прогнозное моделирование: работа с регрессионными моделями больших данных в RStudio.
5. RStudio: даты и временные ряды, загрузка больших данных и тесты на автокорреляцию, качественные переменные, предельные эффекты и ROC кривая. Экспоненциальное сглаживание Холта с корректировкой тренда в моделях временных рядов больших данных в RStudio.
6. Эконометрическое исследование на больших данных Российского мониторинга экономического положения и здоровья населения (РМЭЗ, он же RLMS).
7. Кластерный анализ на больших данных. Анализ потребительской корзины. Использование метода k-средних для сегментирования клиентской базы. Сетевые графы и определение сообществ.
8. Примеры анализа временных рядов больших данных в RStudio. Определение выбросов.

Критерии оценивания:
Максимальная сумма баллов за все лабораторные работы = 55 баллов.
Каждая лабораторная работа №1-3 оценивается максимально в 5 баллов.
Критерии оценки каждой работы:

- 2,5-5 баллов. Задание выполнено в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов; либо задание решено в полном объеме с небольшими погрешностями, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.
- 0-2,4 баллов. Задание выполнено частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с погрешностями либо задание не выполнено или выполнено частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

Каждая лабораторная работа №4-8 оценивается максимально в 8 баллов.
Критерии оценки каждой работы:

- 4,1-8 баллов. Задание выполнено в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов; либо задание решено в полном объеме с небольшими погрешностями, выбраны

верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.

- 0-4 балла. Задание выполнено частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с погрешностями либо задание не выполнено или выполнено частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций

Процедуры оценивания включают в себя текущий контроль и промежуточную аттестацию.

Текущий контроль успеваемости проводится с использованием оценочных средств, представленных в п. 2 данного приложения. Результаты текущего контроля доводятся до сведения студентов до промежуточной аттестации.

Промежуточная аттестация проводится в форме зачета.

Зачет проводится по расписанию промежуточной аттестации в письменном виде. В задании к зачету – 2 теоретических вопроса и 1 задача. Проверка ответов и объявление результатов производится в день зачета. Результаты аттестации заносятся в зачетную ведомость и зачетную книжку студента. Студенты, не прошедшие промежуточную аттестацию по графику сессии, должны ликвидировать задолженность в установленном порядке.

Приложение 2

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Методические указания адресованы студентам очной формы обучения.

Учебным планом предусмотрены следующие виды занятий:

- лекции;
- практические занятия;
- лабораторные работы.

В ходе лекционных занятий рассматриваются источники и способы извлечения, организации хранения и представления данных, теоретические основы статистических методов анализа данных, в том числе Big Data, методология проведения статистического анализа данных, характеризующих финансово-экономическую деятельность бизнес-субъекта, теоретические положения и практические приложения дисциплины «Методы анализа больших данных (Big Data)», даются рекомендации для самостоятельной работы и подготовке к практическим занятиям. Студент должен освоить навыки работы с основными источниками данных, необходимых для финансово-экономического анализа бизнеса.

В ходе практических занятий рассматриваются методы анализа и синтеза в предметной области; современные методы анализа данных; возможные ограничения применения статистических и эконометрических методов; методики совершенствования знаний в области анализа Big Data, даются рекомендации по самостоятельной работе и подготовке к лабораторным работам.

В ходе лабораторных работ углубляются и закрепляются знания студентов по ряду рассмотренных на практических занятиях вопросов, формируются и развиваются навыки использовать современное программное обеспечение (RStudio) для решения экономико-статистических и эконометрических задач обработки данных: построение таблиц, визуализация, проверка гипотез, корреляционно-регрессионный анализ, анализ временных рядов и панельных данных, анализ текстовой и графической информации.

При подготовке к лабораторным работам каждый студент должен:

– изучить рекомендованную учебную литературу;

– подготовить ответы на все вопросы по изучаемой теме.

В процессе подготовки к лабораторным работам студенты могут воспользоваться консультациями преподавателя.

Вопросы, не рассмотренные на практических занятиях и лабораторных работах, должны быть изучены студентами в ходе самостоятельной работы. Контроль самостоятельной работы студентов над учебной программой курса осуществляется в ходе занятий методом устного опроса или посредством тестирования. В ходе самостоятельной работы каждый студент обязан прочитать основную и по возможности дополнительную литературу по изучаемой теме. Выделить непонятные термины, найти их значение в энциклопедических словарях.

Для подготовки к занятиям, текущему контролю и промежуточной аттестации студенты могут воспользоваться электронно-библиотечными системами. Также обучающиеся могут взять на дом необходимую литературу на абонементе университетской библиотеки или воспользоваться читальными залами.

Методические рекомендации по выполнению лабораторных работ

Лабораторная работа №1. Введение в RStudio

Освоить основы работы в RStudio. Работа с электронными таблицами больших данных.

Задание к лабораторной работе №1:

1. Ввод и импортирование данных в RStudio.
2. Подключение библиотек.
3. Работа с массивами данных в RStudio (ввод, чтение, присвоение, вывод данных и т.д.).
4. Работа с электронными таблицами больших данных: ввод данных, фильтрация, сортировка.

Лабораторная работа №2. RStudio: визуализация Big Data, фиктивные переменные, прогнозы, проверка гипотез и ловушка дамми-переменных

Анализ и моделирование рынка труда.

Задание к лабораторной работе №2:

1. Импортировать данные файла wages1 в RStudio, содержащие четыре переменные: exper – стаж работы в годах, male – пол: 1 – для мужчин, 0 – для женщин, school – число лет образования, wage – доход в 1980 году, $/час. Файл содержит наблюдения по 3296 американским индивидам (данные National Longitudinal Survey).
2. Построить гистограмму и плотность для переменной wage.
3. Предварительный анализ данных рекомендуется выполнить самостоятельно. Одним из результатов такого анализа является факт логнормального распределения переменной wage. Поэтому будем использовать новую переменную lnwage, полученную после логарифмирования переменной wage.
4. Подобрать параметры логнормального распределения для wage и вывести для него график функции плотности распределения.
5. Ввести дамми-переменную male – пол: 1 – для мужчин, 0 – для женщин.
6. Посчитать коэффициент корреляции между wage и exper. Построить график корреляционного поля.
7. Постройте регрессионную модель дохода от наиболее значимых факторов. Обоснуйте выбор вида и структуры модели. Оцените ее качество. Сделайте выводы.
8. Сохранить текущий файл R для последующей сдачи.
9. Подготовить отчет.

Лабораторная работа №3. Большие данные и скрипты. Проблемы регрессионного анализа больших данных (мультиколлинеарность и гетероскедастичность) в RStudio

Задание к лабораторной работе №3:

1. Написание скриптов в RStudio.
2. Найти функцию линейной регрессии для простого набора данных house_cost.txt (методом градиентного спуска, аналитически, с помощью функции lm в RStudio).
3. Изучать зависимость средней цены мороженого $price_i$ от количества покупающих мороженого $number_i$ и дамми-переменной на время года $season_i$ (1 - если лето, 0 - если любой другой сезон).
   $price_i=\beta_1+\beta_2 number_i+\beta_3 season_i+\varepsilon_i$
4. Предполагая, что в данных есть гетероскедастичность, принять
   $Var(price_i|X)=\frac{\sigma^2}{number_i}$.
5. Начало матрицы X имеет вид:
$$\begin{pmatrix} 1 & 25 & 1 \\ 1 & 50 & 1 \\ 1 & 42 & 0 \\ \dots & \dots & \dots \end{pmatrix}$$
6. Переделать модель так, чтобы учесть данный вид гетероскедастичности.
7. Чему равен элемент в первой строчке и первом столбце x'_{11} в матрице X', в которой уже есть поправка на вид гетероскедастичности?
8. Подготовить отчет.

Лабораторная работа №4. Прогнозное моделирование: работа с регрессионными моделями больших данных в RStudio

Задание к лабораторной работе №4:

1. Загрузите данные из файла cygage.txt.

2. Используя функцию lm, постройте регрессию, показывающую зависимость возраста исследуемых отложений от глубины залегания, используя веса наблюдений.
3. Постройте прогноз возраста отложений от заданного значения глубины залегания.
4. По тренировочному набору данных titanic_train из пакета library(titanic) с помощью логит-регрессии оценить зависимость вероятности выжить на Титанике от пола Sexi, возраста Agei, стоимости билета Farei и количества мужей/жён и детей на борту Титаника Parchi. Прежде чем оценивать модель, очистите набор данных от пропущенных значений.
5. По данным titanic_train постройте прогноз выживания одинокой женщины 30 лет со средней стоимостью билета.
6. Подготовьте отчет.

Лабораторная работа №5. RStudio: даты и временные ряды, загрузка больших данных и тесты на автокорреляцию, качественные переменные, предельные эффекты и ROC кривая. Экспоненциальное сглаживание Холта с корректировкой тренда в моделях временных рядов больших данных в RStudio

Научиться обрабатывать данные и прогнозировать события, используя возможности логистической регрессии и ROC-анализ.
Задание к лабораторной работе №5:
1. По тренировочному набору данных titanic_train из пакета library(titanic) с помощью логит-регрессии оценить зависимость вероятности выжить на Титанике от пола Sexi, возраста Agei, стоимости билета Farei и количества мужей/жён и детей на борту Титаника Parchi. Прежде чем оценивать модель, очистите набор данных от пропущенных значений.
2.Оценить ограниченную модель, где отсутствует переменная Parchi. Обратите внимание, что сначала надо отобрать массив данных, куда входили переменные расширенной модели, очистить от пропущенных значений. И на этом же массиве оценить ограниченную модель. Провести LR тест на сравнение данных моделей.
3. Создайте обучающую выборку.
4. Проанализируйте полученные данные. Что обычно показано по горизонтали на ROC-кривой в терминах данных по Титанику?
5. Подгрузить временные ряды по ценам акций с помощью команды getSymbols(...). Вытащить данные по компании Гугл (тикер - GOOGL) за период с 1 января 2014 года по 11 мая 2016 года (ресурс - сайт google). Определить чему была равна цена закрытия 9 января 2014 года?
6. Выполнить экспоненциальное сглаживание Холта с корректировкой тренда в моделях данного временного ряда больших данных в RStudio.
7. Создайте отчет.

Лабораторная работа №6. Эконометрическое исследование на больших данных Российского мониторинга экономического положения и здоровья населения (РМЭЗ, он же RLMS)

Проведение исследования, чтобы выявить факторы, влияющие на доход людей, на реальных данных Российского мониторинга экономического положения и здоровья населения (РМЭЗ, он же RLMS).
Задание к лабораторной работе №6:
1. Воспользуйтесь для этого репрезентативной выборкой по индивидам за 2014 год - волна 23 (нужный файл называется r23i_os26a.sav). Для загрузки данных в R воспользуйтесь пакетом rlms.
2. Составьте массив данных с отобранными переменными:
- заработной платой на основе переменной sj13.2;
- возрастом на основе переменной s_age;
- полом на основе переменной sh5;
- наличием высшего образования на основе переменной sj72.5a;
- типом населенного пункта на основе переменной status;
- средней продолжительностью рабочей недели на основе переменной sj6.2;
- семейным положением на основе переменной sj322;
- удовлетворенностью условиями труда на основе переменной sj1.1.2.

3. Отобрать только тех людей, у которых семейное положение входит в данный список:
- никогда в браке не состоял(а);
- состоите в первом зарегистрированном браке;
- состоите в повторном зарегистрированном браке;

- разведены;
- вдовец/вдова.

4. Отобрать только два типа населённого пункта: город и областной центр.
5. Отобрать только две категории степени удовлетворенности условиями труда: полностью удовлетворен и скорее удовлетворен.
6. Отобрать только тех людей, кто на вопрос про высшее образование ответил:
- учились;
- учитесь;
- нет.

7. Из переменной тип населенного пункта сделать дамми-переменную, равную 1 для города и 0 для областного центра.
8. Из переменной удовлетворённость условиями труда сделать дамми-переменную, равную 1 для полностью удовлетворен и 0 для скорее удовлетворен.
9. Из переменной пол сделать дамми-переменную, равную 1 для мужчин и 0 для женщин.
10. Переменную семейное положение необходимо превратить в набор фиктивных переменных. Использовать будем следующие категории:
- никогда в браке не состоял(а);
- состоите в зарегистрированном браке или состоите в повторном зарегистрированном браке;
- разведены;
- вдовец/вдова.

В итоге Вы должны получить 4 фиктивные переменные, отвечающие за принадлежность респондента к одной из этих категорий.
11. Из переменной высшее образование сделать дамми-переменную, равную 1 для тех, кто получил или получает высшее образование, и 0 для тех, кто не получал.
12. В полученном массиве данных должно быть 2523 наблюдения.
13. Создать массив данных, очищенный от пропущенных наблюдений, NA. Таким образом, у Вас должно получиться 2081 наблюдений в массиве без NA! Если оно другое, значит, где-то ошибка и надо пересмотреть предыдущие пункты.
14. Построить регрессионную модель и оценить ее качество. Сформулировать выводы.
15. Сформировать отчет.

Лабораторная работа №7. Кластерный анализ на больших данных. Анализ потребительской корзины. Использование метода k-средних для сегментирования клиентской базы. Сетевые графы и определение сообществ

Научиться осуществлять кластерный анализ больших данных, в том числе используя метод k-средних для сегментирования клиентской базы и метод обработки данных «Самоорганизующиеся карты Кохонена».

Задание к лабораторной работе №7:

1. Выполните необходимые действия по построению карт Кохонена. Проанализируйте результаты, что можно сказать о вероятности возврата кредита для групп 2, 3 и 4?
2. Используя различные отображения карты Кохонена, постройте 3-4 правила выдачи кре-дитов.
3. Ответьте на вопросы: 1) для чего используются карты Кохонена? 2) по какому принципу происходит перенос многомерного пространства на пространство меньшей размерности?
4. Загрузить набор данных iris, доступный по умолчанию в RStudio.
5. Реализовать алгоритм k-means, воспользовавшись функцией mykmeans<-function(data, centroids), и разбейте множество векторов iris[,-5] на 3 кластера. При этом выберите 3 случайных вектора в R^4 и запишите их в centroids.

colors<-mymeans(iris[,-5], centroids)

6. Выведите полученные результаты на графике.
7. Получите лучшее разбиение , выполнив иерархическую кластеризацию методом Варада.
8. Сравните качество кластеризации обоих типов с истинным значением классов iris[,-5].
9. Подготовьте отчет.

Лабораторная работа №8. Примеры анализа временных рядов больших данных в RStudio. Определение выбросов

Задание к лабораторной работе №8:

1. Загрузка данных о котировках фьючерса на индекс РТС в R. Установите библиотеку rusquant.

```
library(rusquant)
getSymbols("SPFB.RTS", from="2007-01-01", src="Finam")
head(SPFB.RTS)
```

2. Возьмите только стоимость фьючерса на момент закрытия биржи.

```
rts<-as.numeric(SPFB.RTS[,4])
plot(rts)
```

3. Весь временной ряд $t_0,t_1,\ldots t_n$ нарежьте на вектора размером w:

$$(t_0;\ldots ; t_{w-1})$$
$$(t_1;\ldots ; t_w)$$
$$(t_2;\ldots ; t_{w+1})$$
$$\ldots$$

4. Для каждого такого вектора определить будет ли расти или падать котировка в момент времени соответствующий w+1 координате каждого вектора. Падение кодировать с помощью -1, а возрастание с помощью 1. Например, пусть $t_w > t_{w-1}$ и $t_{w+1} > t_w$, тогда

$$(t_0;\ldots ; t_{w-1}) \rightarrow 1$$
$$(t_1;\ldots ; t_w) \rightarrow -1$$

Аналогично для всех остальных векторов.

5. Обучите SVM, подберите C и σ, оцените качество предсказания.

6. Подготовить отчет.