МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ
ФГБОУ ВО «Башкирский государственный университет»
Институт экономики, финансов и бизнеса
Институт непрерывного образования

УТВЕРЖДАЮ
Проректор по молодежной политике и международной деятельности

_____ Т.Б. Великжанина
(подпись)
м.п.
«01» 09 2022 г.

СОГЛАСОВАНО
Директор института

_____ К.Е. Гришин
(подпись)

«01» 09 2022 г.

СОГЛАСОВАНО
И.о. директора института непрерывного образования

_____ Е.П. Кислова
(подпись)

«01» 09 2022 г.

ПРОГРАММА ПРОФЕССИОНАЛЬНОЙ ПЕРЕПОДГОТОВКИ

Аналитик данных

(наименование программы)

Специалист по большим данным

(наименование присваиваемой квалификации (в соответствии с профстандартом)

УФА 2022 год

1. ОБЩАЯ ХАРАКТЕРИСТИКА ПРОГРАММЫ

1.1. Цель реализации программы

Программа имеет целью формирование у слушателей таких профессиональных компетенций, как способность разрабатывать продукты на основе встроенной аналитики больших данных и проводить первоначальную обработку первичных статистических данных, необходимых в том числе для выполнения профессиональной деятельности в соответствии со следующими профессиональными стандартами ПС 06.042 «Специалист по большим данным» и ПС 08.022 «Статистик».

1.2. Характеристика нового вида профессиональной деятельности, новой квалификации

а) область профессиональной деятельности в соответствии с ПС 06.042:

Создание и применение технологий больших данных.

б) объекты профессиональной деятельности:

Разработка компьютерного программного обеспечения.

Деятельность консультативная и работы в области компьютерных технологий.

Деятельность, связанная с использованием вычислительной техники и информационных технологий.

Деятельность по обработке данных, предоставление услуг по размещению информации и связанная с этим деятельность.

в) виды и задачи профессиональной деятельности:

Анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры.

Управление этапами жизненного цикла методологической и технологической инфраструктуры анализа больших данных в организации.

г) **уровень квалификации в соответствии с профессиональным стандартом ПС 06.042**: 6 и 7.

1.3. Требование к результатам освоения программы

В результате освоения программы слушатель должен приобрести следующие знания и умения, необходимые для качественного изменения компетенций, указанных в п.1:

знать:

– классификацию типов наборов данных;

– методы сбора и подготовки исходных данных;

– технологии семплирования для получения сбалансированных выборок;

– основные современные методы анализа количественных и факторных данных;

– методы валидации и кросс-валидации при обучении алгоритмов машинного обучения;

– алгоритмы машинного обучения – бинарные модели регрессии, модели регрессии (метод лассо и гребневой регрессии); байесовские классификаторы, алгоритмы бустинга, методы деревьев решений и случайного леса);

– ансамблевые процедуры в машинном обучения;

– алгоритмы обработки входной информации, с применением современных инструментальных средств;

– методы предварительной обработки информации (метод главных компонент, кластерный анализ) для возможности применения продвинутых методов анализа данных;

– метрики качества для оценки алгоритмов машинного обучения;

– методы опорных векторов для решения задач классификации;

– методы подбора нелинейного ядра для решения задач методом опорных векторов;

– методы парсинга данных в сети интернет;

– структуру HTML документа, структуру и параметры HTML тегов, объектную модель DOM;

уметь:

– проводить качественную чистку данных, проводить восполнение данных;

– восполнять выборку до сбалансированного объема;

– проводить процедуры валидации и кросс-валидации для обучения алгоритмов машинного обучения;

– использовать метрики качества для оценки алгоритмов машинного обучения;

– использовать алгоритмы машинного обучения для решения задач классификации;

– использовать алгоритмы машинного обучения для решения задач регрессии;

– проводить ансамблирование алгоритмов машинного обучения для повышения точности решения задач классификации и регрессии;

– применять техники сэмплирования для восполнения баланса выборок, используя средства среды R Studio;

– строить модели на основе алгоритмов машинного обучения для решения задач регрессии и классификации;

– использовать методы машины опорных векторов для задач классификации;

– использовать наивный байесовский классификатор для классификации текстовых сообщений;

– проводить оценку качества алгоритмов машинного обучения;

– ориентироваться в стандарте исключений для роботов сайта;

– извлекать HTML тексты и анализировать их содержимое с помощью библиотек R.

В качестве планируемых результатов освоения программы являются: компетенции в соответствии с трудовыми функциями профессиональных стандартов:

ПК-1 «Способность разрабатывать продукты на основе встроенной аналитики больших данных»;

ПК-2 «Способность проводить первоначальную обработку первичных статистических данных».

1.4. Требования к уровню подготовки поступающего на обучение, необходимому для освоения программы

Лица, желающие освоить дополнительную профессиональную программу, должны иметь или получать среднее профессиональное или высшее образование.

1.5. Трудоемкость обучения

Нормативная трудоемкость обучения по данной программе – 256 часов, включая все виды аудиторной и внеаудиторной (самостоятельной) учебной работы слушателя.

1.6. Форма обучения – очно-заочная с применением электронного обучения и дистанционных образовательных технологий.

1.7. Режим занятий

При данной форме обучения учебная нагрузка устанавливается в соответствие с календарным учебным графиком.

2. СОДЕРЖАНИЕ ПРОГРАММЫ

2.1. Учебный план

Дисциплины	Общая трудоемкость, час	По учебному плану с использованием дистанционных образовательных технологий, час.								СРС час	Промежуточная аттестация (при наличии)
		Контактная работа, час.*				Электронный ресурс, час.					
		всего	из них			всего	из них				
			ЛЗ	ПЗ.	ЛР		ЛЗ	ПЗ	ЛР		
1	2	3	4	5	6	7	8	9	10	11	12
1. Введение в курс	18	6	2		4	6	2		4	6	зачет
2. Основные задачи и работа с данными	20	6	2	2	2	6	2		4	8	зачет
3. Парсинг информации	36	8	2	2	4	20	4	8	8	8	зачет
4. Методы регрессии	36	6	2	2	2	18	4	4	10	12	зачет с оценкой
5. Методы классификации	36	6	2	2	2	18	4	4	10	12	зачет с оценкой
6. Байесовский классификатор	18	4	2	2		8	2	2	4	6	зачет
7. Деревья решений, случайный лес и бустинги	36	6	2	2	2	14	4	4	6	16	зачет
8. Балансирование выборок	18	4	2	2		8	2	2	4	6	зачет
9. Машина опорных векторов	20	4	2	2		8	2	2	4	8	зачет
10. Итоговая аттестация	18									18	итоговый междисципли-нарный экзамен
Итого	256	50	18	16	16	106	26	26	54	100	

2.2. Учебно-тематический план

ПК-1 соответствует ТФ – С/01.8 Разработка продуктов на основе встроенной аналитики больших данных (ПС 06.042 «Специалист по большим данным»)

ПК-2 соответствует ТФ – А/05.4 Первоначальная обработка первичных статистических данных в соответствии с утверждёнными правилами и методиками (ПС 08.022 «Статистик»)

№ п/п	Тема и содержание (краткая аннотация занятия)	Всего часов	В том числе		Самост. работа	Компе-тенция
			Лекции	Практич. и лаборат. занятия		
1	2	3	4	5		
1.	**Введение в курс** Задачи, требующие решения на основе машинного обучения. Постановка задач обучения по прецедентам. Объекты и признаки. Типы шкал: бинарные, номинальные, порядковые, количественные	18	4	8	6	ПК-2
2	**Основные задачи и работа с данными** Работа с данными, преобразование данных из узких таблиц в широкую, фильтрация и объединение данных по определенному признаку. Процедуры подготовки данных для исследований. Упорядоченные и неупорядоченные данные. Транзакционные данные. Определение достаточного количества анализируемых объектов. Верификация. Трансформация. Оптимизация признакового пространства. Разделение выборок, кросс-валидация и метрики качества моделей. Селекция алгоритмов машинного обучения. ROC-анализ. Чувствительность и специфичность. Ложноположительные и ложноотрицательные исходы. Площадь под кривой (Area under curve). Особенности применения ROC-кривых в медицинских исследованиях. Сравнение ROC-кривых между собой. Валидациия и кросс-валидация при обучении алгоритмов машинного обучения	20	4	8	8	ПК-2
3	**Парсинг информации** Парсинг, веб-скрейпинг, краулинг. Вопросы этичности при парсинге. Законодательство и парсинг. API, application programming interface. HTML разметка и модель DOM. Структура web-элемента, узел, тег. Стандарт исключений для роботов, файл robots.txt. Извлечение HTML текстов и анализ их содержимого. Формирование циклов для сбора данных. Установка ограничения на скорость сбора данных согласно Crawl-delay	36	6	22	8	ПК-2
4	**Методы регрессии** Решение задач регрессии, проверка адекватности модели. Метод наименьших квадратов. Проверка адекватности уравнения регресси. Предпосылки для эффективности	36	6	18	12	ПК-1

№ п/п	Тема и содержание (краткая аннотация занятия)	Всего часов	В том числе		Самост. работа	Компе-тенция
			Лекции	Практич. и лаборат. занятия		
	несмещенности и состоятельности оценок. Борьба с мультиколлинеарностью, метод LASSO, гребневая регрессия, выбор параметра регуляризации					
5.	**Методы классификации** Модели множественного выбора с неупорядоченными, бинарными и упорядоченными альтернативами. Условные логит-модель. Вложенные (nested) логит-модели. Оценивание логит-моделей: метод максимального правдоподобия. Интерпретация моделей множественного выбора на основе маржинальных эффектов. Мультиномиальная логит-модель. Примеры моделей с упорядоченными альтернативами. Упорядоченные пробит-модели. Оценка качества для решения задач классификации	36	6	18	12	ПК-1
6.	**Байесовский классификатор** Модели наивного байесовского классификатора. Принцип максимума апостериорной вероятности. Теорема об оптимальности байесовского классификатора. Оценивание плотности распределения: три основных подхода. Непараметрическое оценивание плотности. Вид разделяющей поверхности. Подстановочный алгоритм, его недостатки и способы их устранения. Параметрический наивный байесовский классификатор. Применение НБК для работы с текстом	18	4	8	6	ПК-1
7..	**Деревья решений, случайный лес и бустинги** Алгоритмы дерева принятия решения. Решающее дерево. Жадная нисходящая стратегия «разделяй и властвуй». Алгоритм ID3. Недостатки жадной стратегии и способы их устранения. Проблема переобучения. Редукция решающих деревьев: предредукция и постредукция, прунинг. CARD-деревья. Алгоритмы случайного леса (Random Forest). Алгоритмы случайного леса: косоугольный, синтетический, изолированный и полностью рандомизированный. Определение важности признаков по алгоритмам, основанным на деревьях решений). Ансамблевые методы: бустинги.	36	6	14	16	ПК-1

№ п/п	Тема и содержание (краткая аннотация занятия)	Всего часов	В том числе		Самост. работа	Компе-тенция
			Лекции	Практич. и лаборат. занятия		
	Экстремальный градиентный бустинг. Стохастический градиентный бустинг					
8.	**Балансирование выборок** Технологии сэмплирования. оверсэмплинг, андерсэмплинг, ASMO, SMOTE Проблема неполных данных. Восстановление пропуском. Метод ресамплинга. Цензурирование. Метод исключения некомплектных объектов. Методы с заполнением. Методы взвешивания. Методы, основанные на моделировании	18	4	8	6	ПК-1
9.	**Машина опорных векторов (SVM)** Оптимальная гиперплоскость, зазор между классами, функции потерь для алгоритмов SVM, понятие опорного вектора Функция ядра (kernel functions), спрямляющее пространство, теорема Мерсера. Проверка данных на линейную разделимость, нормализация данных для SVM, Построение линейной решающей функции, предварительно разметка данных, применение алгоритма Вапника. При отсутствии линейного разделения, выбор ядра классификатора: полиномиальное, сигмоидное, радиальное и радиальное Гауссово	20	4	8	8	ПК-1
10.	**Итоговая аттестация** Итоговый междисциплинарный экзамен	18			18	ПК-1 ПК-2

2.3. Дисциплинарное содержание программы

1. Введение в курс (18 ак.часов)

Темы

Задачи, требующие решения на основе машинного обучения. Постановка задач обучения по прецедентам. Объекты и признаки. Типы шкал: бинарные, номинальные, порядковые, количественные

Задания в виде кейса 1 Установка R Studio, определение настроек.

Тест

2. Основные задачи и работа с данными (20 ак. часов)

Темы

Работа с данными, преобразование данных из узких таблиц в широкую, фильтрация и объединение данных по определенному признаку. Процедуры подготовки данных для исследований. Упорядоченные и неупорядоченные данные. Транзакционные данные. Определение достаточного количества анализируемых объектов. Верификация. Трансформация. Оптимизация признакового пространства. Разделение выборок, кросс-валидация и метрики качества моделей. Селекция алгоритмов машинного обучения. ROC-анализ. Чувствительность и специфичность. Ложноположительные и ложноотрицательные

исходы. Площадь под кривой (Area under curve). Особенности применения ROC-кривых в медицинских исследованиях. Сравнение ROC-кривых между собой. Валидациия и кросс-валидация при обучении алгоритмов машинного обучения.

Задания в виде кейса 2 Основные задачи и работа с данными

3. Парсинг Информации (36 ак. часов)

Темы

Парсинг, веб-скрейпинг, краулинг. Вопросы этичности при парсинге. Законодательство и парсинг. API, application programming interface. HTML разметка и модель DOM. Структура web-элемента, узел, тег. Стандарт исключений для роботов, файл robots.txt. Извлечение HTML текстов и анализ их содержимого. Формирование циклов для сбора данных. Установка ограничения на скорость сбора данных согласно Crawl-delay.

Задания в виде кейса 3

Тест

4. Методы регрессии (36 ак. часов)

Темы

Решение задач регрессии, проверка адекватности модели. Метод наименьших квадратов. Проверка адекватности уравнения регресси. Предпосылки для эффективности несмещенности и состоятельности оценок. Борьба с мультиколлинеарностью, метод LASSO, гребневая регрессия, выбор параметра регуляризации.

Задания в виде кейса 4 Построение регрессии, определение параметров регуляризации

5. Методы классификации (36 ак. часов)

Темы

Модели множественного выбора с неупорядоченными, бинарными и упорядоченными альтернативами. Условные логит-модель. Вложенные (nested) логит-модели. Оценивание логит-моделей: метод максимального правдоподобия. Интерпретация моделей множественного выбора на основе маржинальных эффектов. Мультиномиальная логит-модель. Примеры моделей с упорядоченными альтернативами. Упорядоченные пробит-модели. Оценка качества для решения задач классификации.

Задания в виде кейса 5 Построение классификатора на основе моделей бинарной регрессии.

Тест

6. Байесовский классификатор (18 ак. часов)

Темы

Модели наивного байесовского классификатора. Принцип максимума апостериорной вероятности. Теорема об оптимальности байесовского классификатора. Оценивание плотности распределения: три основных подхода. Непараметрическое оценивание плотности. Вид разделяющей поверхности. Подстановочный алгоритм, его недостатки и способы их устранения. Параметрический наивный байесовский классификатор. Применение НБК для работы с текстом.

Задания в виде кейса 6

7. Деревья решений, случайный лес и бустинги (36 ак. часов)

Темы

Алгоритмы дерева принятия решения. Решающее дерево. Жадная нисходящая стратегия «разделяй и властвуй». Алгоритм ID3. Недостатки жадной стратегии и способы их устранения. Проблема переобучения. Редукция решающих деревьев: предредукция и постредукция, прунинг. CARD-деревья. Алгоритмы случайного леса (Random Forest). Алгоритмы случайного леса: косоугольный, синтетический, изолированный и полностью рандомизированный. Определение важности признаков по алгоритмам, основанным на деревьях решений). Ансамблевые методы: бустинги. Экстремальный градиентный бустинг. Стохастический градиентный бустинг

Задания в виде кейса 7

Тест

8. Балансирование выборок (18 ак. часов)

Темы

Технологии сэмплирования. оверсэмплинг, андерсэмплинг, ASMO, SMOTE Проблема неполных данных. Восстановление пропуском. Метод ресамплинга. Цензурирование. Метод исключения некомплектных объектов. Методы с заполнением. Методы взвешивания. Методы, основанные на моделировании.

Задания в виде практического задания (кейс 8)

9. Машина опорных векторов (SVM) (20 ак. часов)

Темы

Что оптимально разделяет гиперплоскость? Как определить зазор между классами? Какие есть функции потерь для алгоритмов? Что такое опорный вектор? Функция ядра (kernel functions), спрямляющее пространство, теорема Мерсера. Как построить конструктивное ядро?

Алгоритм SVM для только для классификации или его можно применять для регрессии? Проверка данных на линейную разделимость, нормализация данных для SVM, Построение линейной решающей функции, предварительно разметка данных, применение алгоритма Вапника. При отсутствии линейного разделения, выбор ядра классификатора: полиномиальное, сигмоидное, радиальное и радиальное Гауссово.

Задания в виде кейса 9.

3. УСЛОВИЯ РЕАЛИЗАЦИИ ПРОГРАММЫ

3.1. Материально-технические условия реализации программы

Наименование специализированных аудиторий, кабинетов, лабораторий	Вид занятий	Наименование оборудования, программного обеспечения
1	2	3
Через платформу ZOOM и с помощью ЭОР на платформе СДО ИНО	лекции	Intel Pentium или аналогичный процессор с тактовой частотой 300MHz и выше. SVGA монитор, с разрешением экрана, как минимум, 800x600 точек и глубиной цвета 16 bit (рекомендуемое разрешение экрана — 1024x768). Звуковая карта, акустическая система или наушники. Доступ в Интернет со скоростью 56 кбит/с и выше. **Программное обеспечение** Операционная система: Windows 7 или более продвинутая, Macintosh, Linux Браузер: Internet Explorer 7 или более продвинутый, Mozilla Firefox (скачать бесплатно: http://www.mozilla.org/download.html) и т.п. **Для просмотра электронных версий учебных курсов необходимо наличие установленных программ:** Microsoft Internet Explorer 7.0 и выше (Загрузить с сайта www.microsoft.com) Adobe Flash Player версии 7.0 и выше (Загрузить с сайта http://www.adobe.com/)
С помощью ЭОР на платформе СДО ИНО	Лабораторные работы	**Аппаратные требования** Intel Pentium или аналогичный процессор с тактовой частотой 300MHz и выше. SVGA монитор, с разрешением экрана, как минимум, 800x600 точек и глубиной цвета 16 bit (рекомендуемое разрешение экрана — 1024x768). Звуковая карта, акустическая система или наушники. Доступ в Интернет со скоростью 56 кбит/с и выше. **Программное обеспечение** Операционная система: Windows 7 или более продвинутая, Macintosh, Linux Браузер: Internet Explorer 7 или более продвинутый, Mozilla Firefox (скачать бесплатно: http://www.mozilla.org/download.html) и т.п. **Для просмотра электронных версий учебных курсов необходимо наличие установленных программ:** Microsoft Internet Explorer 7.0 и выше (Загрузить с сайта www.microsoft.com) Adobe Flash Player версии 7.0 и выше (Загрузить с сайта http://www.adobe.com/) **Для выполнения кейсов требуется установка: R, R Studio, R Tools, Java** **(все открытые ресурсы ПО)**

3.2. Учебно-методическое обеспечение программы

Список основной и дополнительной учебной литературы, электронных ресурсов по программе

Все модули 1-9 сопровождаются разработанными электронными ресурсами, включающими в себя методические указания по выполнению кейса, видео-комментарии к выполнению кейса, скрипт на R для выполнения кейса, шаблон отчёта либо заготовка отчёта в виде markdown-файла. Доступ слушателям курса к электронным ресурсам осуществляется в системе СДО ИНО по ссылке https://sdoino.bashedu.ru/course/view.php?id=249.

1. Анализ данных: учебник для академического бакалавриата / ГУ - Высшая школа экономики; под ред. В. С. Мхитаряна .— Москва : Юрайт, 2016 .— 490 с. (13 экз в библиотеке).
2. Ананьев, В. А. Анализ экспериментальных данных [Электронный ресурс] : учеб. пособие / В. А. Ананьев.— Кемерово : Кемеровский государственный университет, 2009 .— 102 с. [Электронный ресурс] URL=https://biblioclub.ru/index.php?page=book_red&id=232208
3. Чашкин, Ю.Р. Математическая статистика. Анализ и обработка данных : учеб. пособие для студ. высших учеб. заведений .— 2-е изд., перераб. и доп. — Ростов н/Д : Феникс, 2010 .— 236с. . (3 экз в библиотеке).
4. Макшанов, А. В., Журавлев А.Е. Технологии интеллектуального анализа данных: Учебное пособие. — СПб.: Издательство «Лань», 2018 .— 212 с. [Электронный ресурс] URL=https://e.lanbook.com/reader/book/109617/#2.
5. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А.А. Слинкина. – М.: ДМК Пресс, 2015. – 400 с.: ил. [Электронный ресурс] URL=https://e.lanbook.com/reader/book/69955/#4.
6. Чуешев, А. В. Интеграция данных : учебно-методическое пособие / А. В. Чуешев. – Кемерово: Кемеровский государственный университет, 2018. – 281 с.: ил. [Электронный ресурс] URL= https://biblioclub.ru/index.php?page=book&id=495177.

4. ОЦЕНКА КАЧЕСТВА ОСВОЕНИЯ ПРОГРАММЫ

Оценка качества освоения программы включает текущую, промежуточную и итоговую аттестацию обучающихся.

4.1. Критерии оценивания

Оценивание промежуточных результатов обучения

Наименовании модуля	Задание	Балл	Критерии оценки
Введение в курс	Тест	5	Минимальное требование по прохождению теста – 2 балла
	Кейс 1	10	Система R должна быть загружена (минимально допустимый балл - 1 балл)
	Оценка «зачтено», если в сумме задание и тест 3 и более баллов		
Основные задачи и работа с данными	Кейс 2	10	Для минимального балла (3) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов анализа данных
	Оценка «зачтено», если 3 и более баллов		
Парсинг информации	Тест	8	Минимальное требование по прохождению теста – 3 балла
	Кейс 3	9	Для минимального балла (3) должен быть получен правильный ответ при выполнении хотя бы одного задания. Проходной балл - 3
	Оценка «зачтено», если в сумме задание и тест 6 и более баллов		
Методы регрессии	Тест	10	Минимальное требование по прохождению теста – 5 баллов
	Кейс 4	10	Для минимального балла (5) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов моделирования
	Зачет с оценкой: В сумме за задание и тест набрано: 10-11 баллов – **«удовлетворительно»** 12-15 баллов – **«хорошо»** 16 и более баллов – **«отлично»**		
Методы классификации	Тест	10	Минимальное требование по прохождению теста – 4 балла
	Кейс 5	10	Для минимального балла (3) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов моделирования
	Зачет с оценкой: В сумме за задание и тест набрано: 7-10 баллов – **«удовлетворительно»** 11-13 баллов – **«хорошо»** 14 и более баллов – **«отлично»**		

Байесовский классификатор	Кейс 6	5	Для минимального балла (2) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов моделирования
	Оценка «зачтено», если 2 и более баллов		
Деревья решений, случайный лес и бустинги	Кейс 7	20	Для минимального балла (5) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов моделирования
	Оценка «зачтено», если 5 и более баллов		
Балансирова-ние выборок	Тест	8	Минимальное требование по прохождению теста – 3 балла
	Кейс 8	15	Для минимального балла (5) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов моделирования
	Оценка «зачтено», если в сумме задание и тест 8 и более баллов		
Машина опорных векторов	Тест	3	Минимальное требование по прохождению теста – 1 балл
	Кейс 9	10	Для минимального балла (3) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов моделирования
	Оценка «зачтено», если в сумме задание и тест 4 и более баллов		

Оценивание итоговой аттестации (итоговый междисциплинарный экзамен)

Наименовании модуля	Задание	Балл	Критерии оценки
Итоговая аттестация	итоговый тест	12	Проходной балл - 3
	итоговое комплексное задание	30	0-14 – оценка **«неудовлетворительно»** 15-18 – оценка **«удовлетворительно»** 19-23 – оценка **«хорошо»** 24-30 – оценка **«отлично»**
	В случае, если итоговое комплексное задание и итоговый тест оценены: 0-17 баллов – **«неудовлетворительно»** 18-25 баллов - **«удовлетворительно»** 25-35 баллов - **«хорошо»** 35-42 баллов – **«отлично»**		

4.2. Фонд оценочных средств

Кейс-задание 1 (по модулю 1). Установка R Studio

1. Скачать необходимые программы для своей операционной системы

1.1. R • Для win – https://cran.r-project.org/bin/windows/base/

• Для macos – https://cran.r-project.org/bin/macosx/

1.2. RStudio • https://rstudio.com/products/rstudio/download/#download

1.3. Rtools (только для win) • https://cran.r-project.org/bin/windows/Rtools/

2. Установить скаченные программы.

2.1. Порядок установки R -> RStudio -> Rtools

2.2. Для пользователей Windows: • Важно, чтобы путь установки НЕ содержал русских букв

• Нежелательно устанавливать программы в стандартную папку «Program Files», поскольку в дальнейшем возникнут проблемы при установке дополнительных библиотек

• Лучше всего установить R в «C:/R», RStudio в «C:/RStudio», а Rtools в «C:/Rtools»

3. RStudio – это оболочка для языка R. В данном курсе вы будете использовать именно эту программу. После установки запустите RStudio и убедитесь, что она правильно установлена. Также необходимо изменить некоторые стандартные настройки. Для этого откройте раздел Tools – Global options

• 3.1. В разделе General установите настройки на UTF8.:

Кейс-задание 2 (модуль 2). Основные задачи и работа с данными

Цели практического задания

1. Узнать о возможностях R по считыванию данных формата .csv и .xlsx

2. Научиться преобразованию переменных в R: mutate

3. Научиться проводить отбор переменных в R: select

4. Научиться фильтрации наблюдений в R: filter

5. Научиться работать с командами группировки и суммирования в R: group_by и summarise

2 Задание

Порядок выполнения заданий:

1. Ознакомиться с теоретическим материалом по теме (видео-лекциями)

2. Скачать файл с исходным кодом скрипта на языке R (Practice_02.R)

3. Открыть файл скрипта в RStudio. Если комментарии на русском языке отображаются некорректно, то сменить кодировку и перезапустить RStudio (см. видео практики №1)

4. Выполнить команды для подключения пакета tidyverse

5. Провести все действия по скрипту согласно прилагаемому видео

6. В качестве задания создать новый скрипт, который загрузить в качестве выполненного задания

7. Используя встроенный набор ‘cars’ следует выполнить:

• Провести перевод из американских единиц измерения в российские:

– Перевести мили в час (mph) в километры в час (kph) (умножить скорость на 1.61)

– Перевести футы в метры (умножить тормозной путь на 0.31)

– (для пунктов 1 и 2 достаточно обновить существующие переменные и не создавать новые)

• Создайте переменную ratio, которая будет равна тормозной путь (dist) / скорость (speed)

8. Используя встроенный набор ‘swiss’ выполнить Для каждой фильтрации создайте отдельный набор

• наблюдения, в которых доля католиков больше 50% и младенческая смертность меньше 20

• наблюдения, в которых Examination или Education больше 20%

• наблюдения, в которых фертильность больше 60 и младенческая смертность меньше или равна 18

• наблюдения у которых Agriculture принимает значения 1.2 или 7.7 (выполнить через команду %in%)

9. Используя встроенный набор ‘diamonds’ из пакета ggplot2 и выполните

• Сделайте группировку по переменной cut и получите по каждой группе:

– кол-во наблюдений (n()), средняя цена, максимальная цена, медианная каратность

• Сделайте группировку по ДВУМ переменным cut и color получите по каждой группе:

– кол-во наблюдений (n()), медианная цена, минимальная цена, средняя каратность и максимальная каратность

• Загрузить скрипт в личный кабинет

3 О наборах данных для практического занятия

3.1 Набор cars

?cars

View(cars)

В наборе присутствуют показатели:

• speed - скорость в милях в час (mph)

• dist - тормозной путь в футах (ft)

3.2 Набор swiss

?swiss

View(swiss)

Наблюдения по франкоговорящим кантонам Швейцарии за 1888

В наборе присутствуют показатели:

• Fertility - фертильность, общая стандартизированная мера рождаемости

• Agriculture - % мужчин, занятых в сельском хозяйстве

• Examination - % призывников, получивших высшую оценку на армейском экзамене

• Education - % призывников с образованием выше начальной школы

• Catholic - % католиков

• Infant.Mortality - Младенческая смертность. Живорожденные, которые живут менее 1 года.

3.3 Набор diamonds

?diamonds

View(diamonds)

Для выполнения задания вам понадобятся переменные:

• price - цена в долларах

• carat - каратность

• cut - качество огранки алмаза (Fair, Good, Very Good, Premium, Ideal)

• color - цвет алмаза от D (лучшее) to J (худшее)

4 Контрольные вопросы

1. Какие стандарты представления данных позволяет обрабатывать R?

2. Какие команды в R позволяют присоединить к датасету новые переменные ?

3. Какие команды в R позволяют отбрасывать из датасета переменные ?

4. Какие команды в R позволяют отфильтровать из датасета определенные наблюдения?

5. Зачем нужна группировка данных? Возможно ли в R проводить грууппировку одновременно по

двум признакам?

6. Как можно поступать с пропущенными данными?

5 Источники информации

1. https://www.rdocumentation.org/

2. Встроенная справка в RStudio

Результатом выполнения кейс-задания является отчет по кейсу № 2. К отчету предъявляются следующие требования:

1. Четкое формулирование поставленной цели исследования

2. Формулирование задач, решение которых необходимо для достижения поставленной цели.

3. Описание в виде пунктов, тех действий, которые требуются для решения поставленных задач. Все рисунки и таблицы последовательно нумеруются и описываются.

Кейс-задание 3 (модуль 3). Парсинг Информации

Цели практического задания: научиться собирать данные в сети интернет с помощью R

Содержание кейса

1. Анализ файла robots.txt по Стандарту исключений для роботов

2. Извлечение HTML таблицы с помощью библиотеки R: rvest или ralger

3. Получение данных о содержании таблицы

Задание

1. Зайти на сайт http://spatial-economics.com/ и проверить файл robots.txt

Указать какого запрета на сайте нет

Disallow: /administrator/
Disallow: /cache/
Disallow: /components/
Disallow: /images/
Disallow: /archive/

2. Спарсить таблицу со страницы http://spatial-economics.com/en/archive/2020/898-2020-3, в том числе извлечь HTML таблицу и проанализировать ее содержимое с помощью библиотеки R: rvest или ralger Для внесения ответа необходимо открыть полученный результат в R и указать верный ответ.

3. Отметить сколько столбцов и строк в выгруженной таблице

4. Указать какой автор указан в первом столбце второй строке выгруженной таблицы.

Кейс-задание 4 (модуль 4). Построение регрессии, определение параметров регуляризации

1 Содержание кейса

1. Деление данных на обучающую и тестовую выборки
2. Работа с модели линейной регрессии с оценкой коэффициентов МНК
3. Определение их качества (тесты и графики)
4. Работа с Ridge и LASSO моделями
5. Получение прогнозов и метрик

2 Задание

Не забывайте устанавливать параметр set.seed (можете выбирать любое значение)

2.1 Создайте новый скрипт

2.2 Линейные регрессии с оценкой коэффициентов МНК. Используйте данные mtcars

1. Выполните необходимые преобразования:
• vs и am - факторы (factor)
• cyl, gear и carb - упорядоченные факторы (ordered)
2. Разделите выборку на train/test с соотношением 0.8
• Зависимая (целевая) переменная: mpg
3. Модель (1) с одной независимой переменной:
• Независимая переменная: wt
• Посмотрите резюме модели (summary)
• Сделайте прогноз
• Получите метрики качества модели
4. Модель (2) со всеми переменными:
• Независимые переменные: все остальные переменные
• Посмотрите резюме модели (summary)
• Сделайте прогноз
• Получите метрики качества модели
• Сравните метрики моделей (1) и (2)
• Постройте графики распределения ошибок
• Проведите анализ тестов: Бройша-Погана, Бройша-Годфри, Дарбина-Уотсона, Колмагорова-Смирнова, Шапиро-Уилка, VIF

2.3 Дополнительные модели. Используйте данные по Швейцарским кантонам swiss

1. Разделите выборку на train/test с соотношением 0.8:
• Зависимая (целевая) переменная: Fertility
• Преобразуйте наборы для создания моделей Ridge и LASSO
2. Задайте параметр лямбда от 100 до 0 с шагом 0.01
3. Модель Ridge:
• Подберите оптимальную лямбду для Ridge регрессии
• Постройте модель Ridge
• Получите прогноз и метрики качества
4. Модель Lasso:
• Подберите оптимальную лямбду для LASSO регрессии

• Постройте модель LASSO
• Получите прогноз и метрики качества
5. Сравните метрики качества Ridge И LASSO
3 Оформление отчета
Результат этого модуля можно оформить в двух вариантах:
1. Скрипт R: Код + результаты моделей и анализ тестов в комментариях
2. Скрипт R (только код) + Скрины результатов моделей и анализ тестов в Word
Подробнее:
1. Только скрипт R:
• Код всех преобразований
• Результаты оформить через комментарии в скрипте (быстрое создание комментариев Ctrl +
Shift + M)
• summary (значимые переменные и их: коэффициент (Estimate), p-value (Pr(>|t|))), метрики, тесты
• Тесты должны сопровождаться анализом (выполнен/не выполнен, почему)
• В сравнении метрик выпишите какая модель лучше и почему
• Оптимальное значение лямбды (для Ridge и LASSO)
2. Скрипт + Word:
1. Скрипт R:
• код с преобразованиями данных
2. Word:
• Результаты моделей (summary, метрики, графики, тесты) в виде скринов
• К каждому summary выпишите какие переменные оказались значимыми
• Скрины результатов тестов и их анализ (выполнен/не выполнен, почему)
• В сравнении метрик выпишите какая модель лучше и почему
• Оптимальное значение лямбды (для Ridge и LASSO)
4 О наборах данных для практического занятия
4.1 Набор swiss
?swiss
View(swiss)
Наблюдения по франкоговорящим кантонам Швейцарии за 1888
В наборе присутствуют показатели:
• Fertility - фертильность, общая стандартизированная мера рождаемости
• Agriculture - % мужчин, занятых в сельском хозяйстве
• Examination - % призывников, получивших высшую оценку на армейском экзамене
• Education - % призывников с образованием выше начальной школы
• Catholic - % католиков
• Infant.Mortality - Младенческая смертность. Живорожденные, которые живут менее 1 года.
4.2 Набор mtcars
?mtcars
View(mtcars)
В наборе присутствуют показатели:
• mpg - Расход топлива - Miles/(US) gallon
• cyl - Кол-во цилиндров
• disp - Рабочий объем
• hp - Лошадиные силы
• wt - Вес (1000 lbs)
• qsec - 1/4 mile time
• vs - Двигатель (0 = V-shaped, 1 = straight)
• am - Коробка передач (0 = automatic, 1 = manual)
• gear - Количество передних передач
• carb - Количество карбюраторов

Результатом выполнения кейс-задания является отчет по кейсу № 4. К отчету предъявляются следующие требования:

1. Четкое формулирование поставленной цели исследования
2. Формулирование задач, решение которых необходимо для достижения поставленной цели.
3. Описание в виде пунктов, тех действий, которые требуются для решения поставленных задач. Все рисунки и таблицы последовательно нумеруются и описываются. Каждый пункт решения поставленных задач сопровождается анализом принятого решения. При проведении статистических тестов, обязательно выписывается нулевая и альтернативная гипотеза, формулируется принятие решения на обосновано выбранном уровне значимости, указывается критическая область отказа от нулевой гипотезы в пользу альтернативной.
4. В заключении выписывается отобранная адекватная модель с оцененными коэффициентами с указанием под оценками коэффициентов значений t-статистик в скобках или стандартных ошибок коэффициентов. Также приводятся значения маржинальных эффектов и дается их интерпретация.

Кейс-задание 5 (модуль 5). Построение классификатора на основе моделей бинарной регрессии.

1. Провести предварительный анализ исходных данных. Исключить аномальные наблюдения (если такие есть), заполнить пропуски (если они имеются). Провести корреляционный анализ независимых переменных, исключив переменные, значительно коррелирующие с другими переменными (>0,9).
2. Построить статистически значимую модель бинарной регрессии, оценив параметры методом максимального правдоподобия, применяя метод пошагового исключения, в которой все переменные будут статистически значимы. Подобрать функцию распределения, описывающую вероятность положительной альтернативы (например, выживет пациент или умрет) между нормальным распределением (пробит), логистическим (логит) и экстремальным (гомпит) на основе минимума информационных критериев.
3. Проверить качество отобранной модели, подтвердив его значениями коэффициентов R^2 МакФаддена, тестом отношения правдоподобия (LR-тестом), результатами теста Хосмера-Лемешоу и любым тестом на нормальность распределения остатков (например, Колмогорова-Смирнова или Бера-Жарка).
4. Рассчитать маржинальные эффекты и провести интерпретацию коэффициентов модели.
5. Оформить отчет о выполнении задания с приведением условия задачи, результатов решения и выводов.

1 Содержание практики

1. Оформление отчетов в RMarkdown
2. Формат данных .rds
3. Логит, пробит и гомпит модели
4. Селекция моделей (AIC, BIC)
5. Коэффициенты, графики и тесты
6. Маржинальные эффекты
7. Прогнозирование и матрицы сопряженности

2 Задание

Вам предстоит определить какие показатели влияют на удовлетворённость жизнью

2.1 Создайте новый скрипт RMarkdown

1. Загрузите файл Joy.rds
2. Разделите выборки на train/test с соотношением 0.8
3. Зависимая (целевая) переменная: Joy

2.2 Моделирование

Вам необходимо построить логит, пробит и гомпит модели. Для каждого типа вам необходимо повторить следующие шаги:

1. Постройте модель на всех переменных

2. Проверьте какие переменные оказались не значимыми

3. Перестройте модель без эти переменных

4. Повторяйте действия пока все переменные не будут значимыми

Примечание: У переменной Health несколько уровней: “Плохое”, “Среднее”, “Хорошее”. Если хотя бы один из уровней окажется значимым, переменную удалять нельзя

2.3 Селекция и тесты

1. С помощью информационных критериев (AIC, BIC) выберите лучшую модель

2. Тесты и коэффициенты: коэффициент детерминации Макфаддена, Likelihood-ratio, Колмогоров- Смирнов, Хосмер-Лемешоу, графики распределения ошибок

3. Проведите анализ тестов

2.4 Прогнозирование и маржинальные эффекты

1. Получите прогноз

2. Сформируйте матрицу сопряженности

3. Посчитайте маржинальные эффекты

Примечание: В пакете mfx нет маржинальных эффектов для гомпит-модели, если лучшей моделью окажется гомпит-модель, получите маржинальные эффекты для логит или пробит модели. В зависимости от того, какая из моделей будет лучше по AIC и BIC

3 Оформление отчета

Отчет необходимо оформить в RMarkdown. Отключите warning и message в куске кода (чанке) с подключением пакетов. Скрывать код (echo) из отчета не нужно.

4 О наборе данных для практического занятия

Источник данных: «Российский мониторинг экономического положения и здоровья населения НИУ ВШЭ (RLMS-HSE)», проводимый НИУ Высшая школа экономики и «Демоскоп» при участии Института социологии РАН. (Сайты обследования RLMS-HSE: hse.ru/rlms). Мониторинг представляет собой серию общенациональных репрезентативных опросов, проводимых на базе вероятностной стратифицированной многоступенчатой территориальной выборки, разработанной при участии ведущих мировых экспертов в этой области.

В этой практике вы будете использовать данные 23 волны

Целевая переменная:

• fac - Joy - Вы удовлетворены своей жизнью в целом в настоящее время?

Независимые переменные:

• fac - Sex - Пол респондента

• num - Age - Количество полных лет

• fac - Job - Вы удовлетворены или не удовлетворены вашей работой в целом?

• fac - Money - Вы удовлетворены своим материальным положением в настоящее время?

• num - Children - Сколько всего у вас детей?

• fac - Marriage - Состоите ли вы в зарегистрированном браке?

• fac - Health - Как Вы оцениваете ваше здоровье?

• fac - Chronic - Есть ли у вас другие хронические заболевания?

• fac - Smoking - Вы курите в настоящее время?

• fac - Rise - Вы удовлетворены или не удовлетворены возможностями для вашего профессионального роста?

• fac - Vacation - В течение последних 12 месяцев вы были в оплачиваемом отпуске?

• fac - Decrease - В течение последних 12 месяцев вы перешли на более низкую должность?

Результатом выполнения кейс-задания является отчет по кейсу № 5. К отчету предъявляются следующие требования:

1. Четкое формулирование поставленной цели исследования

2. Формулирование задач, решение которых необходимо для достижения поставленной цели.

3. Описание в виде пунктов, тех действий, которые требуются для решения поставленных задач. Все рисунки и таблицы последовательно нумеруются и описываются. Каждый пункт решения поставленных задач сопровождается анализом принятого решения. При проведении статистических тестов, обязательно выписывается нулевая и альтернативная гипотеза, формулируется принятие решения на обосновано выбранном уровне значимости, указывается критическая область отказа от нулевой гипотезы в пользу альтернативной.

4. В заключении выписывается отобранная адекватная модель с оцененными коэффициентами с указанием под оценками коэффициентов значений t-статистик в скобках или стандартных ошибок коэффициентов. Также приводятся значения маржинальных эффектов и дается их интерпретация.

Кейс-задание 6. (Модуль 6). Фильтрация на основе наивного байесовского классификатора

Провести классификацию IT-приложений для решения подобных задач согласно наивному байесовскому классификатору, предварительно проверив валидность исходных данных. Сформировать правила классификации, сочетаемые с классом априорных вероятностей по данным тестовой выборки. Пересчитать на основе оцененной модели классификации апостериорные вероятности для полученных ранее правил на основе данных обучающей выборки. Вывести матрицу неточности для каждой зависимой переменной. Проверить гипотезу о корректности сформированной модели байесовской классификации. Построить график проведения байесовской классификации на обучающей выборке. Построить график предсказания, апостериорной вероятности. Построить сценарный прогноз отнесения IT-приложений к определенным классам, используя найденную модель.

Результатом выполнения задания является отчет по кейсу 6. К отчету предъявляются следующие требования:

1. Формулирование задачи, решение которых необходимо в ходе выполнения лабораторной работы (например, провести классификацию программного обеспечения автоматизирования технологических процессов от уровня проектирования до внедрения).

2. Описание данных для тестовой и обучающей выборок, заключение о валидности данных.

3. Четко сформулированные выводы по результатам выполнения лабораторной работы.

4. Оценить качество классификатора на тестовой выборке. Все графики и таблицы должны иметь сквозную нумерацию.

Кейс-задание 7 (модуль 7). Алгоритмы дерева принятия решения, алгоритмы случайного леса, бустинги

Провести классификацию объектов, для решения используя классификатор деревьев решения и алгоритм randomforest, предварительно проверив валидность исходных данных. Выбрать признак и значения порога, по которому происходит оптимальное по заданному критерию разбиение в алгоритме randomforest. Задать максимальное число объектов в вершине-листа дерева, для определения критерия останова алгоритма. Построить алгоритм на размеченных данных. Провести классификацию объектов, для решения используя классификатор метод бустинга, предварительно проверив валидность исходных данных. Применить алгоритм градиентного бустинга, используя правило жадного наращивания. Найти пары наиболее оптимальных параметров, где под оптимизацией следует понимать принцип явной максимизации отступов, минимизировать функционал ошибки. Провести классификацию объекто, используя алгоритм бустинга AdaBoost с экспоненциальной функцией потерь

1 Содержание практики

1. Пакет caret

• Кросс-валидация

• Гридсерч

• Построение моделей

• Метрики

2. Модель дерева решений
3. Модель случайного леса
4. Модель полностью случайного леса
5. Модель XGBoost

2 Задание

В практическом задании вы будете использовать классический набор по титанику. Вам необходимо предсказать какие пассажиры переживут крушение, а какие погибнут.

2.1 Создайте новый скрипт RMarkdown.

1. Загрузите файл titanic.rds
2. Разделите выборки на train/test с соотношением 0.8
3. Зависимая (целевая) переменная: Survived

2.2 Моделирование

Для каждой модели получите прогноз и матрицу сопряженности. Также для моделей построенных через caret выпишите какие гиперпараметры оказались наилучшими.

1. Сформируйте параметры для кросс-валидации:
• метод: cv
• число фолдов: 6
2. Модель дерева решений
3. Модель случайного леса:
• стандартный способ (функция randomForest): ntree 150, mtry = 2
• через caret. Сетка: mtry = 1, 2, 3, 4
4. Модель полностью случайного леса. Сетка: mtry = 1, 2, 3, 4; numRandomCuts = 1, 2, 3, 4
5. Модель XGBoost
6. Подведите итоги какая модель оказалась наилучшей

3 Оформление отчета

Отчет необходимо оформить в RMarkdown. Отключите warning и message в куске кода (чанке) с подключением пакетов. Скрывать код (echo) из отчета не нужно.

4 О наборе данных для практического занятия

Классический набор машинного обучения. Судьба пассижиров Титаника.

Целевая переменная:
• fac - Survived - Выжил ли пассажир (1 - да, 0 - нет)

Независимые переменные:
• fac - Pclass - Каким классом плыл пассажир (1, 2, 3)
• fac - Sex - Пол
• num - Age - Возраст
• num - SibSp - Число братьев, сестер и супругов на борту корабля
• num - Parch - Число родителей и детей на борту

Результатом выполнения задания является отчет по кейсу 7. К отчету предъявляются следующие требования:

1. Формулирование задачи, решение которых необходимо в ходе выполнения лабораторной работы (например, провести классификацию программного обеспечения автоматизирования технологических процессов от уровня проектирования до внедрения).
2. Описание данных для тестовой и обучающей выборок, заключение о валидности данных.
3. Обосновать выбор признака и значения порога, по которому происходит оптимальное по заданному критерию разбиение в алгоритме randomforest
4. Применить алгоритм градиентного бустинга, используя правило жадного наращивания.
5. Найти пары наиболее оптимальных параметров, где под оптимизацией следует понимать принцип явной максимизации отступов, минимизировать функционал ошибки.
6. Провести классификацию объектов, для решения используя классификатор - метод экстремального бустингна с корректно подобранными весами.

7. Оценить качество классификаторов на тестовой выборке. Все графики и таблицы должны иметь сквозную нумерацию.

8. Четко сформулированные выводы по результатам выполнения кейса.

Кейс-задание 8. Технологии сэмплирования

Определить в исходной информации количество случаев мажоритарного и миноритарного класса. Применить алгоритмы одностороннего сэмплирования. Применить способ повышения количества образцов миноритарного класса – метод SMOTE (Synthetic Minority Oversampling Technique). Выбрать лучшей алгоритм сэмплирования. Применить метод адаптивного искусственного увеличения числа примеров миноритарного класса ASMO (Adaptive Synthetic Minority Oversampling). Выбрать лучшей алгоритм сэмплирования.

Результатом выполнения задания является отчет по работе № 8. К отчету предъявляются следующие требования:

1. Описать исходные данные.
2. Применить алгоритм сэмплирования SMOTE.
3. Применить алгоритм сэмплирования ASMO.

Кейс 9. Построение модели классификации с помощью алгоритма машины опорных векторов (SVM).

1. Скачать данные creditcard.csv в формате csv, и скрипт R.
2. Установить пакеты dplyr, caret, caTools, e1071, kernlab, затем их подключить с помощью команды library
3. Считать данные creditcard.csv и разделить их на обучающую и тестовую выборку
4. Рассчитать модель классификации с помощью пакета kernlab, в режиме масштабирования (scaled=TRUE) и без него, при этом перебрать три варианта ядра:

"rbfdot" - радиальная базисная функция

"polydot" - полиномиальная базисная функция

"vanilladot" - линейная базисная функция

5. Определить матрицу неточности для всех вариантов моделей.
6. Рассчитать модель классификации с помощью пакета caret, указав гиперпараметр method = "svmRadial" (алгоритм SVM с радиальной базисной функцией)
7. Определить матрицу неточности для всех вариантов моделей
8. Сделать выводы о качестве полученных моделей классификации. Оформить отчёт.

Примечание: В кейсе дано задание, данные creditcard.csv и два скрипта на R (оригинальный и распараллеленный). Параллельная версия кейса по SVM будет использовать 4 процессорных ядра. Тесты показали, что на 4-х ядерном процессоре можно ускориться как минимум в 3 раза относительно последовательнго расчета. Если раскомментировать "cl <- makeCluster(detectCores())", будут задействованы все имеющиеся ядра (логические процессоры).

Оценка качества освоения программы осуществляется в виде защиты проекта по итоговому комплексному заданию, позволяющему оценить сформированность профессиональных компетенций ПК-1 и ПК-2.

До выполнения итогового комплексного задания допускаются лица, имеющие сданные на положительные оценки кейсы 1-9 и получившие допустимый балл по тестам промежуточной аттестации (критерии оценки кейсов и тестов сведены в таблицу в пункте 4).

Слушатель считается аттестованным, если полностью справился с заданием, предоставил отчёт в виде html-файла, разработанного средствами markdown, смог защитить выполненный проект и провести корректную интерпретацию полученных результатов.

Пример итогового комплексного задания

Цели задания: оценка сформированности компетенции ПК-1 и ПК-2

1 В итоговом задании необходимо использовать данные переписи США. Данные предоставляются. Целевая переменная: доход респондента выше или ниже 50000$ в год.

1.1 Создайте новый скрипт RMarkdown.

1. Загрузите датасет Practice_08_dataset.rds

2. Разделите выборки на train/test с соотношением 0.7

3. Зависимая (целевая) переменная: Target

1.2 Моделирование

1. Вам необходимо построить три модели:

a. Стандартная модель (без caret, сэмплирования или кросс-валидаций)

b. Модель с сэмплированием (оверсэмплинг или SMOTE)

c. Модель с сэмплированием (оверсэмплинг или SMOTE) и кросс-валидацией (метод: cv; число фолдов: 5)

• Рекомендуемые алгоритмы: дерево, случайный лес, полностью случайный лес, XGBoost.

• Можно использовать один и тот же алгоритм не больше двух раз

2. Для каждой модели необходимо получить прогноз и построить матрицу сопряженности

3. Подведите итоги какая модель оказалась наилучшей

2 Оформление отчета

Отчет необходимо оформить в RMarkdown и скомпилировать в HTML.

Отключите warning и message в куске кода (чанке) с подключением пакетов. Скрывать код (echo) из отчета не нужно.

3 О наборе данных для итогового задания

Данные переписи США за 1995 год

Целевая переменная:

• fac - Target - Бинарная переменная годовой доход: больше 50000$/год или меньше 50000$/год

Независимые переменные:

• num - Age - Возраст
• fac - Workclass - Сфера занятости
• fac - Education - Образование
• num - Education_num - Образование, число лет
• fac - Martial_status - Семейное положение
• fac - Occupation - Профессия
• fac - Race - Раса
• fac - Sex - Пол
• num - Hours_Per_Week - Сколько часов в неделю работает
• fac - Native_Country - Родная страна

5. КАЛЕНДАРНЫЙ УЧЕБНЫЙ ПЛАН-ГРАФИК ОСВОЕНИЯ ПРОГРАММЫ

Наименование дисциплины	1 нед	2 нед	3 нед	4 нед	5 нед	6 нед	7 нед	8 нед	Итого часов
1. Введение в курс	18								18
2. Основные задачи и работа с данными	12	8							20
3. Парсинг информации		26	10						36
4. Методы регрессии			22	14					36
5. Методы классификации				18	18				36
6. Байесовский классификатор					14	4			18
7. Деревья решений, случайный лес и бустинги						28	8		36
8. Балансирование выборок							18		18
9. Машина опорных векторов							6	14	20
Итоговая аттестация								18	18
Итого часов	30	34	32	32	32	32	32	32	256

6. СОСТАВИТЕЛИ ПРОГРАММЫ

Лакман Ирина Александровна, канд. техн. наук, доцент
Тимирьянова Венера Маратовна, канд. экон. наук, доцент

Руководитель программы:

Фамилия, имя, отчество (при наличии)	Место основной работы и должность, ученая степень и ученое звание	Ссылка на веб-страницы с портфолио	Фото в формате jpeg
Лакман Ирина Александровна	ФГБОУ ВО «Башкирский государственный университет», заведующая лабораторией исследования социально-экономических проблем регионов, канд. техн. наук, доцент	http://scivision.pro/lackman.html	

Руководитель программы _____ И.А. Лакман
подпись