I. ОБРАЗОВАТЕЛЬНАЯ ПРОГРАММА

Федеральное государственное бюджетное образовательное учреждение высшего образования
«БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

УТВЕРЖДАЮ

Ректор

_____ Н.Д. Морозкин

«5» 10 2020 г.

Дополнительная профессиональная программа повышения квалификации

«Анализ данных на основе машинного обучения»

72 час.

СОГЛАСОВАНО

Директор ИНО _____ Т.Б.Великжанина

«5» 10 2020г

УФА 2020

ОБЩАЯ ХАРАКТЕРИСТИКА ПРОГРАММЫ

1. **Цель программы** дать систематизированное представление о современных подходах к анализу данных средствами машинного обучения, познакомить с основными принципами и этапами решения задач регрессии и классификации, научить навыкам применения технологий машинного обучения, в том числе ансамблированию алгоритмов, проверке качества алгоритмов с помощью процедур валидации и кросс-валидации, научить слушателей курса техникам сэмлирования в случае отсутствия сбалансированности классов во входной информации, проводить обучение алгоритмов машинного обучения в современных информационных средах (R Studio), в конечном итоге сформировать на базовом уровне компетенцию компетенции цифровой экономики: Управление информацией и данными.

2. **Планируемые результаты обучения:**
   a. Знание (осведомленность в областях)

2.1.1 классификации типов наборов данных,
2.1.2 методов сбора и подготовки исходных данных,
2.1.3 технологий семплирования для получения сбалансированных выборок,
2.1.4 основных современных методов анализа количественных и факторных данных (технология one-hot, частотный анализ,);
2.1.5 методов валидации и кросс-валидации при обучении алгоритмов машинного обучения;
2.1.6 алгоритмов машинного обучения – бинарные модели регрессии, модели регресии (метод лассо и гребневой регрессии); байесовские классификаторы, методы опорных векторов, алгоритмы бустинга, методы деревьев решений и случайного леса;
2.1.7 ансамблевых процедур в машинном обучения.
2.1.8 методы предварительной обработки информации для возможности применения продвинутых методов анализа данных,
2.1.9 метрик качества для оценки алгоритмов машинного обучения.

2.2 Умение (способность к деятельности):

2.2.1 проводить качественную чистку данных, проводить восполнение данных;
2.2.2 восполнять выборку до сбалансированного объема;
2.2.3 проводить процедуры валидации и кросс-валидации для обучения алгоритмов машинного обучения;
2.2.4 использовать метрики качества для оценки алгоритмов машинного обучения;
2.2.5 использовать алгоритмы машинного обучения для решения задач классификации;
2.2.6 использовать алгоритмы машинного обучения для решения задач регрессии;
2.2.7 проводить ансамблирование алгоритмов машинного обучения для повышения точности решения задач классификации и регрессии

2.3 Навык (использование конкретных инструментов):

2.3.1 владеть техниками сэмплирования для восполнения баланса выборок, используя средства среды R Studio;

2.3.2 навыками построения моделей на основе алгоритмов машинного обучения для решения задач регрессии и классификации.

2.3.3 навыками оценки качества алгоритмов машинного обучения

3 **Требования к слушателям** (возможно заполнение не всех полей)

3.1 Образование: высшее, средне-специальное

3.2 Квалификация: инженер, математик

3.3 Наличию опыта профессиональной деятельности: работа в Excel.

3.4 Предварительное освоение иных дисциплин/курсов /модулей: высшая математика, теория вероятностей и математическая статистика, общая теория статистики

4.Учебный план программы «Анализ данных на основе машинного обучения»

№ п/п	Модуль	Всего, час	Виды учебных занятий		
			лекции	практические занятия	самостоятельная работа
0	**Входное тестирование**	1		1	
1	**Модуль 1. Введение в курс**	5	2	2	**1**
2	**Модуль 2. Основные задачи и работа с данными**	8	2	3	**3**
3	**Модуль 3. Методы регрессии**	11	3	4	**4**
4	**Модуль 4. Методы классификации**	11	3	4	**4**
5	**Модуль 5. Байесовский классификатор**	9	2	3	**4**
6	**Модуль 6. Деревья решений, случайный лес и бустинги**	10	2	4	**4**
7	**Модуль 7. Балансирование выборок**	9	2	3	**4**
Итоговая аттестация			**Зачёт**		
Итоговое задание		8	Зачет - Кейс-проект		

5.Календарный план-график реализации образовательной программы

(дата начала обучения – дата завершения обучения) в текущем календарном году, указания на периодичность набора групп (не менее 1 группы в месяц)

№ п/п	Наименование учебных модулей	Трудоёмкость (час)	Сроки обучения
1	**Модуль 1 – Введение в курс**	5	1.11.2020
2	**Модуль 2 – Основные задачи и работа с данными**	8	2.11.2020- 3.11.2020-
3	**Модуль 3 – Методы регрессии**	11	4.11.2020- 7.11.2020

4	**Модуль 4 – Методы классификации**	11	8.11.2020-10.11.2020
5	**Модуль 5 – Байесовский классификатор**	9	11.11.2020-13.11.2020
6	**Модуль 6 – Деревья решений, случайный лес и бустинги**	10	14.11.2020-16.11.2020
7	**Модуль 7 – Балансирование выборок**	9	17.11.2020-18.11.2020
	Итоговое задание	8	19.11.2020-21.11.2020
Всего:		72	1.11.2020-21.11.2020

6.Учебно-тематический план программы « Анализ данных на основе машинного обучения»

№ п/п	Модуль / Тема	Всего, час	Виды учебных занятий			Формы контроля
			лекции	практические занятия	самостоятельная работа	
0	**Входное тестирование**	**1**			**1**	**Тест**
1	**Модуль 1: Введение в курс**	5	2	2	1	Кейс 1
2	**Модуль 2: Основные задачи и работа с данными**	8	2	3	3	
2.1	Работа с данными, преобразование данных	4	1	2	1	Кейс 2
2.2	Разделение выборок, кросс-валидация и метрики качества моделей	4	1	1	2	Кейс 2
3	**Модуль 3 – Методы регрессии**	11	3	4	**4**	
3.1.	Оценка уравнений регрессии	6	2	2	2	Тест
3.2.	Метод Lasso и гребневая регрессия	5	1	2	2	Кейс 3
4	**Модуль 4 – Методы классификации**	11	3	4	4	
4.1.	Логистическая регрессия	6	2	2	2	Кейс 4
4.2.	Маржинальные эффекты	5	1	2	2	Тест
5	**Модуль 5 - Байесовский классификатор**	9	2	3	**4**	

5.1	Предобработка текста	5	1	2	2	Кейс 5
5.2	Байесовская классификация	4	1	1	2	Кейс 5
6	**Модуль 6. Деревья решений, случайный лес и бустинги**	10	2	4	**4**	
6.1	Деревья решений	4	1	2	1	Кейс 6
6.2	Методы ансамблирования	6	1	2	3	
7	**Модуль 7. Балансирование выборок**	9	2	3	**4**	Кейс 7
8	Итоговая аттестация	8			8	Кейс-итоговый проект

7. Учебная (рабочая) программа повышения квалификации «Анализ данных на основе машинного обучения»

7.1 Модуль 1 «**Введение в курс**» (5 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**.**

7.2 Модуль 2 «**Основные задачи и работа с данными**» (8 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**

7.3 Модуль 3 **Методы регрессии** (11 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 3 Построение регрессии, определение параметров регуляризации

7.4 Модуль 4 **Методы классификации** (11 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 4 Построение классификатора на основе моделей бинарной регрессии.

7.5 Модуль 5 **Байесовский классификатор** (9 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 5

7.6 Модуль 6 **Деревья решений, случайный лес и бустинги** (10 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 6

7.7 Модуль 7 **Балансирование выборок** (9 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.

Описание практико-ориентированных заданий и кейсов

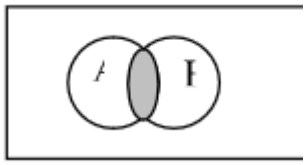
	Номер темы/модуля	**Наименование практического занятия**	**Описание**
1.1	1	Кейс-задание 1. **Установка R Studio**	Цель практического задания: формирование навыков работы с R Studio, установка соответствующих настроек

1.2.	2.	Кейс-задание 2. **Основные задачи и работа с данными**	Цели практического задания: научится работать с данными различной природы, проводить преобразование данных
1.3	2	Кейс-задание 3. Построение регрессии, определение параметров регуляризации	Цели практического задания: оценка методом МНК уравнения регрессии, проверка его адекватности. Проведение регуляризации для метода ЛАССо и гребневой регресси.
1.4	2	Кейс-задание 4 (модуль 4). **Построение классификатора на основе моделей бинарной регрессии**	Цели практического задания: решение задач классификации методами логистической регрессии, определение матрицы сопряжённости и ROC-анализа. Интерпретация результатов на основе маржинальных эффектов.
1.5	3	**Кейс-задание 5. (Модуль 5). Фильтрация на основе наивного байесовского классификатора**	Цели практического задания: проведение классификации предобработанного текста на основе наивного байесовского классификатора
1.6.	4	**Кейс-задание 6. (модуль 6). Алгоритмы дерева принятия решения, алгоритмы случайного леса, бустинги**	Цели практического задания: решение задачи классификации с помощью деревьев решения, ансамблирование алгоритмов машинного обучения за счет случайного леса и бустинга.
1.7	4	Кейс-задание 7. **. Технологии сэмплирования**	Цели практического задания: переход к сбалансированным выборкам с помошью андесэмлинговых и оверсэмлинговых процедур
1.8	5	Кейс-задание 8. Пространственно-ограниченная кластеризация	Цели практического задания: оценка сформированности компетенции на базовом уровне

8.Оценочные материалы по образовательной программе

8.1. Вопросы тестирования по модулям

№ модуля	Вопросы входного тестирования	Вопросы промежуточного тестирования	Вопросы итогового тестирования
0	1. Сумма двух событий – это событие, состоящее A. в одновременном их появлении B. в появлении по крайней мере одного из них C. в их последовательном появлении D. в не появлении ни одного из них 2. Множество результатов, отобранных из генеральной совокупности, называют A. Выборкой B. Вариационным рядом C. Статистикой критерия D. Точечными оценками 3. Статистическая гипотеза – это утверждение о свойствах A. генеральной совокупности B. выборки C. конкретного объекта 4. Значение признака, находящееся в середине вариационного ряда наблюдений, a. мода; b. средняя арифметическая; c. медиана; d. частота; e. частость. 5. Названия гипотезы, противоположной проверяемой: A. нулевая B. простая C. конкурирующая D. альтернативная E. Сложная 6. Интервал возможных значений парного коэффициента корреляции при наличии между величинами X и Y отрицательной, но не функциональной связи: A. (–1; 0) B. (0; 1) C. (–1; -0,5) D. (–0,5; 0) E. [–1; 0] 7. Алгебраическая квадратная матрица является вырожденной, если: A. Определитель матрицы равен нулю		

	B. Ранг матрицы равен размерности матрицы C. У нее имеется обратная матрица D. Определитель матрицы равен единице 8. Заданы множества A={2,3,4,5} и D={3,4,5}. Верным для них будет утверждение: A. Множество A - подмножество множества D B. Множество D - подмножество множества A C. Множество A и множество D равны D. Множество A - множество-степень множества D 9. На рисунке показано:  A. A∪ B B. A ∩B C. A∈ B D. A/ B 10. Для того, чтобы два вектора были ортогональны необходимо и достаточно, чтобы A. их скалярное произведение равнялось нулю B. их векторное произведение равнялось нулевом вектору C. их векторное произведение равнялось нулю **D.** их скалярное произведение равнялось нулевому вектору		
2		1. Несмещенная оценка – это оценка A. с нулевым математическим ожиданием B. с математическим ожиданием, равным истинному значению параметра C. с минимально возможной дисперсией D. сходящаяся по вероятности к истинному значению параметра **3.** Линейная регрессионная модель: A. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + \varepsilon$ B. $y = \beta_0 + \beta_1 x + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$	

		C. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \varepsilon$ D. $y = \beta_0 \beta_1 x \beta_2 x^2 x^k + \varepsilon$ **4.** Регрессоры, коэффициенты при которых значимы: $\tilde{y}$ =-12,5+6,07x_1-2,15x_2+23,8x_3 (7,3) (6,2) (1,3) (3,3) (t-статистики) A. отсутствуют *B.* x_1 C. x_2 D. x_3 **5.** Увеличение регрессора x_1 на единицу своего измерения в уравнении $\tilde{y}$ =-10,5+6,57x_1-0,22x_2+7,8x_3 (7,3) (5,7) (4,3) (12,3) (t-статистики) приводит к следующему изменению среднего значения зависимой переменной A. росту на 6,57 единиц своего измерения B. уменьшению на 6,57 единиц своего измерения C. росту на 6,57% D. уменьшению на 6,57% E. уменьшению на 10,5% i. Явления, вызывающие необходимость изменения функциональной формы регрессии: A. ненулевое значение математического ожидания остатков B. нулевое значение математического ожидания остатков C. постоянство дисперсии остатков для всех наблюдений D. автокорреляция остатков 6 Гомоскедастичность ошибок в регрессионных моделях означает, что они имеют A. одинаковую дисперсию для всех наблюдений B. изменяющуюся дисперсию с ростом значений регрессоров C. одинаковое математическое ожидание для всех наблюдений D. изменяющееся математическое ожидание с ростом значений регрессоров 7 Факторы, включаемые в множественную регрессионную модель, должны удовлетворять следующим условиям: A. быть количественно измеримыми	

		B. каждый фактор должен быть достаточно тесно связан с результатом C. факторы не должны быть мультиколлинеарными D. факторы должны быть мультиколлинеарными 8. В классической регрессионной модели факторы $x_{1i}, x_{2i}, ..., x_{ji}, ..., x_{mi}$ являются величинами A. случайными B. детерминированными C. постоянными D. гетероскедастичными 9. Последствия мультиколлинеарности: A. оценки параметров становятся ненадежными B. небольшое изменение исходных данных приводит к существенному изменению оценок параметров модели C. изменение исходных данных не влияет на оценки параметров модели D. невозможно определить влияние каждого фактора на результативный показатель 10 Интервал возможных значений коэффициента при произвольном регрессоре A. (–1; 1) B. (0; 1) **C.** $(-\infty;\ \infty)$ D. (–1; 0)	
3		1. Укажите какие модели используют в случае качественной зависимой переменной: A. линейную регрессионную модель B. логит-модель C. пробит модель D. обратную модель E. функцию Энгеля 2. Неизвестные коэффициенты в логит - модели находятся с помощью A. Метода наименьших квадратов B. Метода максимального правдоподобия C. Обобщённого метода наименьших квадратов D. Рекурсивного метода 3 Критерий согласия Хосмера-Лемешоу проводится: A. Для оценки качества бинарной модели регрессии B. Для интерпретации полученных результатов моделирования	

		C. Для оценки согласованности расчетных и фактических данных D. для тестирования остатков на нормальность распределения 4.Интерпретация результатов моделирования с помощью бинарной регрессии проводится на основе: A. на основе приростного анализа B. на основе анализа коэффициентов эластичности C. на основе расчета средних значений фактора D. на основе маржинальных эффектов. 5 Признаками модели выбора неупорядоченных альтернатив является: A. наличие латентной переменной, связывающей альтернативы B. неслучайный (упорядоченный) уровень полезности выбора альтернативы C. случайный уровень полезности выбора альтернативы D. возможность разбиения модели на систему бинарных моделей. 6 Для модели выбора упорядоченных альтернатив имеет место следующие утверждения: A. Сумма маржинальных эффектов по одному фактору по каждой альтернативе равна 0 B. .Сумма маржинальных эффектов по всем факторам по одной альтернативе равна 0 C. Сумма маржинальных эффектов по одному фактору по каждой альтернативе равна 100 % D. Сумма маржинальных эффектов по всем факторам по одной альтернативе равна 100% 7 Суть метода кросс-валидации заключается: 1. проведение кросс-проверки качества модели на различных выборках 2. В разбиении исходной выборки на пять частей и обучении модели на четырех из них, а на одной тестирование результатов 3. В разбиении обучающей выборки на несколько частей и обучении модели на четырех из них, а на одной тестирование результатов 4. В делении выборки на обучающую и тестовую 8. Под чувствительностью модели понимают: 1. Доля истинно положительных примеров	

		2. Число верно классифицированных положительных примеров 3. Число положительных примеров, классифицированных как отрицательные 4. Доля положительных примеров, классифицированных как отрицательные 9.На основе ROC-кривой можно сделать вывод: 1. О качестве классификатора 2. О влиянии критерия отсечения на качество модели классификации 3. О точности предсказания класса 4. О наличии пропусков в исходных данных 10 Выберите один или несколько ответов. При оценке качества мультклассовой классификации: 1. Рассчитывается общая матрица сопряженности и по ней находится среднее метрик качества для каждого класса 2. Рассчитывается общая матрица сопряженности и по ней находится средневзвешенное метрик качества для каждого класса 3. Рассчитываются отдельно для каждого класса матрицы сопряженности и по ним находятся среднее метрик качества для каждого класса **4.** Рассчитываются отдельно для каждого класса матрицы сопряженности и по ним находятся средневзвешенное метрик качества для каждого класса	

8.2. описание показателей и критериев оценивания, шкалы оценивания

Минимальным проходным баллом теста считается 60% верных ответов по результатам суммарно 2 попыток

Кейс в 10 баллов: Максимум 10 баллов - выставляется при выполнении всех требований к отчету, подробном описании всех этапов и представлении выводов, увязывающих выполненный кейс с встречающимися в практике задачами.
8-10 баллов выставляется при выполнении всех требований к отчету, подробном описании всех этапов или представлении выводов, увязывающих выполненный кейс с встречающимися в практике задачами
6-7 баллов выставляется при выполнении всех требований к отчету, подробном описании отдельных этапов или кратких выводов.
Минимально допустимый балл 5 баллов -выставляется при выполнении минимального требования к отчету кейса

Кейс в 20 баллов: Максимум 20 баллов - выставляется при выполнении всех требований к отчету, подробном описании всех этапов и представлении выводов, увязывающих выполненный кейс с встречающимися в практике задачами.

16-20 баллов выставляется при выполнении всех требований к отчету, подробном описании всех этапов или представлении выводов, увязывающих выполненный кейс с встречающимися в практике задачами

11-15 баллов выставляется при выполнении всех требований к отчету, подробном описании отдельных этапов или кратких выводов.

Минимально допустимый балл 10 баллов -выставляется при выполнении минимального требования к отчёту кейса**.**

8.3. итоговое задание по всей образовательной программе

Цели задания: оценка сформированности компетенции по анализу данных и машинному обучению (способность управлять информацией и знаниями) на базовом уровне.

Итоговое комплексное Задание
В итоговом задании необходимо использовать данные переписи США. 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:
1. Стандартная модель (без caret, сэмплирования или кросс-валидаций)
2. Модель с сэмплированием (оверсэмплинг или SMOTE)
3. 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 или Word. 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

8.4. тесты и обучающие задачи (кейсы), иные практикоориентированные формы заданий

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 комнаду %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). 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 № 3. К отчету предъявляются следующие требования:

4. Четкое формулирование поставленной цели исследования
5. Формулирование задач, решение которых необходимо для достижения поставленной цели.
6. 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.

7. 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 4 (модуль 4). **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 ока-
жется 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 № 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). 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 5. 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 6. (модуль 6). 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 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. 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 7. 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 № 7. К отчету предъявляются следующие требования:

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

8.5. описание процедуры оценивания результатов обучения

Наименовании модуля	Задание	Балл	Критерии оценки
Входное тестирование	ТЕСТ	10	Проходной балл -5
Модуль 1 – Введение в курс	Кейс 1	10	Система R должна быть загружена (минимально допустимый балл - 8 баллов)
Модуль 2 – Основные задачи и работа с данными	Кейс 2	10	Для минимального балла (5) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов анализа данных
Модуль 3 – Методы регрессии	Кейс 3	10	Для минимального балла (5) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов моделирования
	Тест	10	Проходной балл - 6
Модуль 4 – Методы классификации	Кейс 4	20	Для минимального балла (10) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов моделирования
	Тест	10	Проходной балл - 6
Модуль 5 – Байесовский классификатор	Кейс 5	5	Для минимального балла (3) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов моделирования
Модуль 6 – Деревья решений, случайный лес и бустинги	Кейс 6	20	Для минимального балла (10) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов моделирования
Модуль 7 – Балансирование выборок	Кейс 7	10	Для минимального балла (5) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов моделирования
Итоговая аттестация (итоговый проект)	Задание на проект	40	Итоговое задание считается выполненным на минимально допустимый балл 20, если есть полностью сформированное задание, но

			отсутствует интерпретация полученных результатов
	Минимальный балл для получения зачета по КПК - 78, максимальный - 145		

9.Организационно-педагогические условия реализации программы

9.1. Кадровое обеспечение программы

№ п/п	Фамилия, имя, отчество (при наличии)	Место основной работы и должность, ученая степень и ученое звание (при наличии)	Ссылки на веб-страницы с портфолио (при наличии)	Фото в формате jpeg	Отметка о полученном согласии на обработку персональных данных
1	Лакман Ирина Александровна	ФГБОУ ВО «Башкирский государственный университет", заведующая лабораторией исследования социально-экономических проблем регионов, к.т.н., доцент		**Загружено на платформу**	Да

9.2.Учебно-методическое обеспечение и информационное сопровождение

Учебно-методические материалы	
Методы, формы и технологии	Методические разработки, материалы курса, учебная литература
Методы организации учебно-познавательной деятельности: практический; Форма: дистанционная; Технологии: Информационно – коммуникационная технология; Кейс технология	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. Барский, А.Б. Логические нейронные сети : учебное пособие.— Москва : Интернет-Университет Информационных Технологий, 2007.— 352 с. [Электронный ресурс] URL=https://biblioclub.ru/index.php?page=book_red&id=232983 **Дополнительная литература** 5. 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 6. Сидняев, Н.И, Вилисова Н.Т.Введение в теорию планирования эксперимента: учеб. пособие.— М. : МГТУ им. Баумана. Золотая коллекция, 2011. – 463 с. [Электронный ресурс] URL=https://e.lanbook.com/book/106359#authors 7. Бонцанини М. Анализ социальных медиа на Phyton / пер. с англ. А.В. Логунова.— М : ДМК Пресс, 2018.— 288 с.: ил. . [Электронный ресурс] URL=https://e.lanbook.com/reader/book/108129/#4 8. 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

Информационное сопровождение	
Электронные образовательные ресурсы	Электронные информационные ресурсы
http://sdo.bashedu.ru/course/view.php?id=2267	https://www.kaggle.com/

9.3.Материально-технические условия реализации программы

Вид занятий	Наименование оборудования, программного обеспечения
Лекции, практические занятия	**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)