ОЦЕНОЧНЫЕ МАТЕРИАЛЫ
ПО ДИСЦИПЛИНЕ

Машинное обучение и компьютерный анализ данных

Код модуля
1150603(1)

Модуль
Модели прикладной экономики

Екатеринбург

Оценочные материалы составлены автором(ами):

№ п/п	Фамилия, имя, отчество	Ученая степень, ученое звание	Должность	Подразделение
1	Кобылкин Константин Сергеевич	кандидат физико-математических наук, без ученого звания	Доцент	экономики

Согласовано:

Управление образовательных программ И.Ю. Русакова

Авторы:

- **Кобылкин Константин Сергеевич**, **Доцент**, **экономики**

1. СТРУКТУРА И ОБЪЕМ ДИСЦИПЛИНЫ Машинное обучение и компьютерный анализ данных

1.	**Объем дисциплины в зачетных единицах**	5	
2.	**Виды аудиторных занятий**	Лекции Лабораторные занятия	
3.	**Промежуточная аттестация**	Зачет	
4.	**Текущая аттестация**	Домашняя работа	2

2. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ (ИНДИКАТОРЫ) ПО ДИСЦИПЛИНЕ МОДУЛЯ Машинное обучение и компьютерный анализ данных

Индикатор – это признак / сигнал/ маркер, который показывает, на каком уровне обучающийся должен освоить результаты обучения и их предъявление должно подтвердить факт освоения предметного содержания данной дисциплины, указанного в табл. 1.3 РПМ-РПД.

Таблица 1

Код и наименование компетенции	**Планируемые результаты обучения (индикаторы)**	**Контрольно-оценочные средства для оценивания достижения результата обучения по дисциплине**
1	**2**	**3**
ПК-9 -Способен производить расчеты для оценки эконометрических моделей на основе применения прикладных программ (Прикладная и международная экономика)	П-1 - Владеть навыками организации эмпирического анализа, использования прикладного программного обеспечения для решения возникающих задач У-2 - Уметь строить модели и оценивать их эффективность с использованием современного программного обеспечения для решения экономико-статистических и эконометрических задач	Домашняя работа № 1 Домашняя работа № 2 Зачет Лабораторные занятия Лекции
ПК-10 -Способен составлять прогноз социально-экономических и финансовых показателей деятельности систем	У-1 - Уметь применять современный математический инструментарий и программное обеспечение для решения экономико-статистических и эконометрических задач	Домашняя работа № 1 Домашняя работа № 2 Зачет Лабораторные занятия Лекции

разного уровня, разрабатывать и применять проектные решения с учетом фактора неопределенности (Прикладная и международная экономика)		

3. ПРОЦЕДУРЫ КОНТРОЛЯ И ОЦЕНИВАНИЯ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ В РАМКАХ ТЕКУЩЕЙ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ МОДУЛЯ В БАЛЬНО-РЕЙТИНГОВОЙ СИСТЕМЕ (ТЕХНОЛОГИЧЕСКАЯ КАРТА БРС)

3.1. Процедуры текущей и промежуточной аттестации по дисциплине

1. Лекции: коэффициент значимости совокупных результатов лекционных занятий – 0.5		
Текущая аттестация на лекциях	**Сроки – семестр, учебная неделя**	**Максимальная оценка в баллах**
домашняя работа №1	4	*50*
домашняя работа №2	7	*50*
Весовой коэффициент значимости результатов текущей аттестации по лекциям – 0.6		
Промежуточная аттестация по лекциям – зачет **Весовой коэффициент значимости результатов промежуточной аттестации по лекциям – 0.4**		
2. Практические/семинарские занятия: коэффициент значимости совокупных результатов практических/семинарских занятий – не предусмотрено		
Текущая аттестация на практических/семинарских занятиях	**Сроки – семестр, учебная неделя**	**Максимальная оценка в баллах**
Весовой коэффициент значимости результатов текущей аттестации по практическим/семинарским занятиям– не предусмотрено		
Промежуточная аттестация по практическим/семинарским занятиям–нет **Весовой коэффициент значимости результатов промежуточной аттестации по практическим/семинарским занятиям– не предусмотрено**		
3. Лабораторные занятия: коэффициент значимости совокупных результатов лабораторных занятий –0.5		
Текущая аттестация на лабораторных занятиях	**Сроки – семестр, учебная неделя**	**Максимальная оценка в баллах**
Работа в ходе лабораторных занятий	6	100
Весовой коэффициент значимости результатов текущей аттестации по лабораторным занятиям -не предусмотрено		
Промежуточная аттестация по лабораторным занятиям –нет		

Весовой коэффициент значимости результатов промежуточной аттестации по лабораторным занятиям – не предусмотрено		
4. Онлайн-занятия: коэффициент значимости совокупных результатов онлайн-занятий –не предусмотрено		
Текущая аттестация на онлайн-занятиях	**Сроки – семестр, учебная неделя**	**Максимальная оценка в баллах**
Весовой коэффициент значимости результатов текущей аттестации по онлайн-занятиям -не предусмотрено		
Промежуточная аттестация по онлайн-занятиям –нет **Весовой коэффициент значимости результатов промежуточной аттестации по онлайн-занятиям – не предусмотрено**		

3.2. Процедуры текущей и промежуточной аттестации курсовой работы/проекта

Текущая аттестация выполнения курсовой работы/проекта	**Сроки − семестр, учебная неделя**	**Максимальная оценка в баллах**
Весовой коэффициент текущей аттестации выполнения курсовой работы/проекта– не предусмотрено		
Весовой коэффициент промежуточной аттестации выполнения курсовой работы/проекта– защиты – не предусмотрено		

4. КРИТЕРИИ И УРОВНИ ОЦЕНИВАНИЯ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ МОДУЛЯ

4.1. В рамках БРС применяются утвержденные на кафедре/институте критерии (признаки) оценивания достижений студентов по дисциплине модуля (табл. 4) в рамках контрольно-оценочных мероприятий на соответствие указанным в табл.1 результатам обучения (индикаторам).

Таблица 4

Критерии оценивания учебных достижений обучающихся

Результаты обучения	**Критерии оценивания учебных достижений, обучающихся на соответствие результатам обучения/индикаторам**
Знания	Студент демонстрирует знания и понимание в области изучения на уровне указанных индикаторов и необходимые для продолжения обучения и/или выполнения трудовых функций и действий, связанных с профессиональной деятельностью.
Умения	Студент может применять свои знания и понимание в контекстах, представленных в оценочных заданиях, демонстрирует освоение умений на уровне указанных индикаторов и необходимых для продолжения обучения и/или выполнения трудовых функций и действий, связанных с профессиональной деятельностью.
Опыт /владение	Студент демонстрирует опыт в области изучения на уровне указанных индикаторов.
Другие результаты	Студент демонстрирует ответственность в освоении результатов обучения на уровне запланированных индикаторов. Студент способен выносить суждения, делать оценки и формулировать выводы в области изучения.

	Студент может сообщать преподавателю и коллегам своего уровня собственное понимание и умения в области изучения.

4.2 Для оценивания уровня выполнения критериев (уровня достижений обучающихся при проведении контрольно-оценочных мероприятий по дисциплине модуля) используется универсальная шкала (табл. 5).

Таблица 5

Шкала оценивания достижения результатов обучения (индикаторов) по уровням

Характеристика уровней достижения результатов обучения (индикаторов)				
№ п/п	**Содержание уровня выполнения критерия оценивания результатов обучения (выполненное оценочное задание)**	**Шкала оценивания**		
		Традиционная характеристика уровня		**Качественная характеристика уровня**
1.	Результаты обучения (индикаторы) достигнуты в полном объеме, замечаний нет	Отлично (80-100 баллов)	Зачтено	Высокий (В)
2.	Результаты обучения (индикаторы) в целом достигнуты, имеются замечания, которые не требуют обязательного устранения	Хорошо (60-79 баллов)		Средний (С)
3.	Результаты обучения (индикаторы) достигнуты не в полной мере, есть замечания	Удовлетворительно (40-59 баллов)		Пороговый (П)
4.	Освоение результатов обучения не соответствует индикаторам, имеются существенные ошибки и замечания, требуется доработка	Неудовлетворительно (менее 40 баллов)	Не зачтено	Недостаточный (Н)
5.	Результат обучения не достигнут, задание не выполнено	Недостаточно свидетельств для оценивания		Нет результата

5. СОДЕРЖАНИЕ КОНТРОЛЬНО-ОЦЕНОЧНЫХ МЕРОПРИЯТИЙ ПО ДИСЦИПЛИНЕ МОДУЛЯ

5.1. Описание аудиторных контрольно-оценочных мероприятий по дисциплине модуля

5.1.1. Лекции

Самостоятельное изучение теоретического материала по темам/разделам лекций в соответствии с содержанием дисциплины (п. 1.2. РПД)

5.1.2. Лабораторные занятия

Примерный перечень тем

1. Основные задачи и понятия машинного обучения. Применения к решению бизнес-задач. Примеры некоторых метрик качества прогноза

2. Работа с библиотеками в среде Anaconda.

3. Линейные методы регрессии, классификации и прогнозирования

4. Метод опорных векторов для линейно неразделимых данных. Ядерный метод опорных векторов. Напоминание понятия о скользящем контроле (кроссвалидации). Подбор гиперпараметров в методе опорных векторов

5. Пример решения задачи многоклассовой классификации с помощью метода опорных векторов. Метод опорных векторов для задач регрессии

6. Регуляризация в линейных моделях. Особенности обработки данных в линейных моделях (нормализация данных, кодировка нечисловых признаков). Калибровка прогнозов. Сравнение качества различных линейных моделей.

7. Деревья регрессии и классификации. Принципы вычисления прогноза в деревьях, общие черты и отличия от линейных моделей. Процесс обучения деревьев и критерии построения ветвлений. Подбор критерия ветвления под конкретную метрику качества. Понятие о стратифицированном скользящем контроле.

8. Подбор гиперпараметров деревьев. Информативность признаков. Понятие о композиции деревьев.

9. Случайные леса как сложные композиции деревьев. Принципы обучения и формирования прогноза в случайных лесах. Обсуждение факторов, влияющих на качество прогноза случайных лесов.

10. Случайные леса: продвинутые аспекты настройки. Градиентный бустинг: основные принципы обучения и формирования прогноза. Основные гиперпараметры градиентного бустинга.

11. Особенности настройки и обучения градиентного бустинга. Библиотека Optuna для эффективного перебора гиперпараметров бустинга. AutoML для табличных данных

LMS-платформа − не предусмотрена

5.2. Описание внеаудиторных контрольно-оценочных мероприятий и средств текущего контроля по дисциплине модуля

Разноуровневое (дифференцированное) обучение.

Базовый

5.2.1. Домашняя работа № 1

Примерный перечень тем

1. Метод опорных векторов для линейно неразделимых данных. Ядерный метод опорных векторов. Напоминание понятия о скользящем контроле (кроссвалидации). Подбор гиперпараметров в методе опорных векторов

Примерные задания

Задание 1

1) В разделе **Кодировка категориальных и порядковых признаков** из переменной `categorical_features` убрать добавление генетических категориальных признаков `features_dict['genetic']['category']`, запустить вычисления ниже, вычислить предсказания на тестовой выборке;

2) загрузить предсказания на [сайт](https://www.kaggle.com/c/competition-2-yandex-shad-spring-2020/overview) соревнования в разделе Late submission;

3) посмотреть свою ошибку на тесте (Private Score) в разделе My submissions. Записать ее в ноутбуке."

Задание 2

Оставаясь в условиях предыдущего задания, т.е. отбросив генетические категориальные признаки попробовать:\n",

1) разбить данные по обучающей (new_data и y_data) и тестовой выборке (new_data_test) на две части по признаку `V298` (см. раздел **Разбиение выборки на части по признаку V298**)

2) обучить на каждой подвыборке обучающей выборки отдельную модель логистической регрессии, подобрать для каждой гиперпараметр C

3) вычислить на обеих подвыборках тестовой выборки предсказания;

4) объединить предсказания на тестовых выборках (`np.concatenate((predictions_1, predictions_2))`, `np.concatenate((np.array(test_index)[np.array(new_data_1.index)], np.array(test_index)[np.array(new_data_2.index)]))`)

5) повторить пункты 2-3 предыдущего задания.

Задание 3 Попробовать:

1) на полных (не разбитых данных) в разделе **Создание данных для обучения модели** вычесть из данных по признакам-вероятностям `data[features_dict['medical']['probability']]` среднее или медиану (использовать либо контекстную функцию .mean(axis=0), т.е. `data[features_dict['medical']['probability']].mean(axis=0)`, либо применить функцию библиотеки `numpy`, вычисляющую медиану `np.median(data[features_dict['medical']['probability']], axis=0)`)

2) на полученных данных обучить модель логистической регрессии, а также ядерной логистической регрессии на основе функции Nystrom как в предыдущей лабораторной; подобрать C

3) выполнить пункты 3-5 из предыдущего раздела.

Задание 4 Попробовать:

1) с помощью функции `sklearn.impute.SimpleImputer` в разделе **Заполнение пропусков** заполнить пропуски наиболее частым значением

2) на полученных данных обучить логистическую регрессию, подобрать C

3) выполнить пункты 3-5 первого задания.

Задание 5

Проверить на нормальность и логнормальность распределения всех числовых и вероятностных признаков (см. прилагаемый ноутбук с визуализацией для числовых признаков). Применить функцию `scipy.stats.shapiro`

LMS-платформа − не предусмотрена

5.2.2. Домашняя работа № 2

Примерный перечень тем

1. Деревья регрессии и классификации. Принципы вычисления прогноза в деревьях, общие черты и отличия от линейных моделей. Процесс обучения деревьев и критерии построения ветвлений. Подбор критерия ветвления под конкретную метрику качества. Понятие о стратифицированном скользящем контроле.

2. Подбор гиперпараметров деревьев. Информативность признаков. Понятие о композиции деревьев.

Примерные задания

По известным медицинским показателям требуется определить, разовьётся ли у пациента диабет в ближайшие 5 лет.

Известно, что пациенты были здоровы в момент первого обследования. Пациенты, у которых диабет был обнаружен в первые 2 года исследования, также не вошли в выборку. Пациенты сдавали анализ на глюкозу не реже раза в год. Если у пациента на очередном обследовании был обнаружен диабет, его участие в исследовании на этом заканчивалось. Если пациент отказывался от продолжения участия в исследовании, считалось, что он остался здоров.

Функционал ошибки - LogLoss, т.е. минус логправдоподобие предсказаний.

Стратегия перебора гиперпараметров случайного леса такая:

- для быстроты, вначале при небольшом числе деревьев (`n_estimators`) подбираются все остальные гиперпараметры, в первую очередь `max_features`, во вторую - все остальные параметры (можно вместе, что медленно, можно по одному), среди которых наиболее важный `min_samples_leaf`;

- затем подбирается параметр `n_estimators`; обычно, чем больше `n_estimators`, тем выше качество, поэтому `n_estimators` выбирается достаточно большим, чтобы обеспечить приемлемое качество, но не слишком большим, чтобы скорость обучения была приемлемой.

Задание

- добавить в гридсерч перебор гиперпараметра `min_samples_leaf`, взять значения 2, 100, 1000, ...;

- задать разные `random_state` в параметрах `RandomForestClassifier` и усреднить прогнозы нескольких случайных лесов с разными `random_state` (см. [документацию](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html) на `RandomForestClassifier`);

- добавить несколько взаимодействий признаков, например, сочетание нескольких факторов риска: (глюкоза при первом визите >=6 и человек курит) или (глюкоза при первом визите >=6 и человек злоупотребляет алкоголем) или (глюкоза при первом визите >=6 и у человека есть родственники диабетики) или (глюкоза при первом визите >=6 и возраст > 65). Попробовать запустить модель случайного леса с добавленными признаками;

- обучить на признаках логистическую регрессию и сравнить со случайным лесом на кросс-валидации.

Для каждого эксперимента сгенерировать предсказания на тесте и отправить в разделе `Late Submission` на [соревновании](https://www.kaggle.com/c/competition-2-yandex-shad-spring-2021/overview). Написать в ноутбуке полученные ошибки в public и private лидерборде.

LMS-платформа − не предусмотрена

5.3. Описание контрольно-оценочных мероприятий промежуточного контроля по дисциплине модуля

5.3.1. Зачет

Список примерных вопросов

1. Задачи классификации

2. Задачи регрессии и прогнозирования

3. Кластерный анализ

4. Приведите пример применения машинного обучения при решении бизнес-задач

5. Опишите основные метрики качества прогноза

6. Опишите общую концепцию линейных моделей

7. На основе каких принципов формируется прогноз в линейных моделях?

8. Каковы основные идеи метода опорных векторов в случае, когда выборки из классов линейно разделимы.

9. Охарактеризуйте понятие отступа как меры уверенности в разделении классов.

10. Каковы особенности метода опорных векторов со штрафом за неправильную классификацию объектов.

11. Ядерный метод опорных векторов, границы его применимости.

12. Полный цикл применения метода опорных векторов для задачи классификации изображений.

13. Метод главных компонент.

14. Проиллюстрируйте процесс оценки прогнозного качества метода и перебор гиперпараметров моделей.

15. Идеи применения метода опорных векторов для задач регрессии.

16. Основные принципы регуляризации регрессионных коэффициентов в линейных моделях

17. Каковы способы калибровки прогнозов моделей с целью получения вероятности принадлежности к классу в качестве прогноза

18. Основные принципы вычисления прогноза в деревьях

19. Дайте понятие критерием построения ветвлений

20. Поясните связь между критерием ветвления и метрикой качества прогноза.

21. Поясните понятие стратифицированного скользящего контроля и особенностей его применения для несмещенной оценки качества прогноза в задачах классификации

22. Дайте определение случайного леса, опишите основные принципы обучения таких моделей и формирования их прогнозов

23. Метод бутстрепа и метод случайных подпространств.

24. Основные гиперпараметры случайных лесов, влияющие на их прогнозное качество (размер случайного подпространства, число усредняемых деревьев).

LMS-платформа − не предусмотрена

5.4 Содержание контрольно-оценочных мероприятий по направлениям воспитательной деятельности

Направления воспитательной деятельности сопрягаются со всеми результатами обучения компетенций по образовательной программе, их освоение обеспечивается содержанием всех дисциплин модулей.