1. **Паспорт образовательной программы**

**«Практический анализ данных»**

|  |  |
| --- | --- |
| **Версия программы** | 01 |
| **Дата версии** | 08.10.2020 г. |

1. **Сведения о провайдере:**

|  |  |  |
| --- | --- | --- |
| **№** |  | **Содержание** |
| **1.1** | Провайдер | Общество с ограниченной ответственностью «Юниум» |
| **1.2** | Логотип образовательной организации |  |
| **1.3** | Провайдер, ИНН | 7720364967 |
| **1.4** | Ответственный за программу, ФИО | Кондратьев Дмитрий Сергеевич |
| **1.5** | Ответственный за программу, должность | Генеральный директор |
| **1.6** | Ответственный за программу, телефон | +7 (903) 503-46-46 |
| **1.7** | Ответственный за программу,  Е-mail | dk@unium.ru |

1. **Основные данные:**

|  |  |  |
| --- | --- | --- |
| **№** |  | **Содержание** |
| **2.1** | Название программы | «Практический анализ данных» |
| **2.2** | Ссылка на страницу программы | http://unium.dataschool.digital/ |
| **2.3** | Формат обучения | онлайн |
|  | Подтверждение от ОО наличия возможности реализации образовательной программы с применением электронного обучения и (или) дистанционных образовательных технологий с возможностью передачи данных в форме элементов цифрового следа | Образовательная организация подтверждает возможность реализации программы с применением электронного обучения и (или) дистанционных образовательных технологий. Для реализации образовательной программы используется платформа «Zoom». |
| **2.4** | Уровень сложности | Базовый |
| **2.6** | Количество академических часов | 72 ак. часа |
|  | Практикоориентированный характер образовательной программы: не менее 50 % трудоёмкости учебной деятельности отведено практическим занятиям и (или) выполнению практических заданий в режиме самостоятельной работы (кол-во академических часов) | Образовательная программа имеет практикоориентированный характер, 41 ак. час трудоемкости образовательной программы отведено выполнению практических заданий в режиме самостоятельной работы, а также весь лекционный материал нацелен на освоение практических знаний и навыков. |
| **2.6** | Стоимость обучения одного обучающегося по образовательной программе, а также предоставление ссылок на 3 (три) аналогичные образовательные программы иных организаций, осуществляющих обучение, для оценки объективности стоимости или обоснование уникальности представленной образовательной программы в случае отсутствия аналогичных образовательных программ на рынке образовательных услуг | Стоимость обучения одного слушателя составляет  30 000 рублей.  Ссылки на аналогичные образовательные программы:  <https://netology.ru/programs/analytics-for-executives>  <https://newprolab.com/ru/bigdata/>  <https://www.bigdataschool.ru>  <https://bigdatateam.org/ru>  <https://www.cdo.training> |
| **2.7** | Минимальное количество человек на курсе | 100 |
| **2.8** | Максимальное количество человек на курсе | 2 000 |
| **2.9** | Данные о количестве слушателей, ранее успешно прошедших обучение по образовательной программе | Более 500 человек |
| **2.10** | Формы аттестации | Тестирование и состязание на Kaggle |
|  | Указание на область реализации компетенций цифровой экономики, к которой в большей степени относится образовательная программа, в соответствии с Перечнем областей | Большие данные |

1. **Аннотация программы:**

Образовательная программа по практическому анализу данных – это возможность получить навыки по решению реальных бизнес-задач с помощью инструментов машинного обучения и работы с большими данными. В рамках курса вы узнаете, как с помощью современных методов и библиотек для анализа данных (OCR (компьютерное зрение), NLP (анализ языка), стек библиотек по работе с BigData и др.) решать задачи бизнеса. Программа ориентирована на достижение бизнес-результата, поэтому с первых занятий будут рассмотрены типичные аналитические задачи из разных индустрий, и, прежде чем перейти к изучению современных методов их решения, важно будет научиться понимать бизнес-смысл задачи, экономический эффект и какие инструменты применять для решения задач в зависимости от их специфики. Каждое занятие сопровождается практическими примерами и заданиями.

Участники, успешно завершившие курс, смогут самостоятельно писать основные алгоритмы машинного обучения, включая подготовку данных, отбор и формирование признаков, выбор оптимальных библиотек и их применение, а также выбор оптимальных метрик качества с учетом бизнес-задачи. Программа рассчитана на слушателей, имеющих знания в области линейной алгебры и теории вероятности на уровне первого курса технического вуза.

По результатам обучения у участников будет возможность поучаствовать в реальном состязании на Kaggle.

УТВЕРЖДАЮ

Генеральный директор

ООО «Юниум»

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Д.С. Кондратьев

«8» октября 2020 г

1. ОБРАЗОВАТЕЛЬНАЯ ПРОГРАММА (ДПО)

Общество с ограниченной ответственностью «Юниум»

УТВЕРЖДАЮ

Генеральный директор

ООО «Юниум»

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Д.С. Кондратьев

«8» октября 2020 г

**ОСНОВНАЯ ПРОФЕССИОНАЛЬНАЯ ОБРАЗОВАТЕЛЬНАЯ ПРОГРАММА ДОПОЛНИТЕЛЬНОЙ ПРОФЕССИОНАЛЬНОЙ ПОДГОТОВКИ**

**(повышение квалификации)**

**по направлению подготовки**

**«Практический анализ данных»**

72 ак. часа

Форма обучения: онлайн

г. Москва, 2020 г.

**ОБЩАЯ ХАРАКТЕРИСТИКА ПРОГРАММЫ**

**Аннотация программы:**

Образовательная программа по практическому анализу данных – это возможность получить навыки по решению реальных бизнес-задач с помощью инструментов машинного обучения и работы с большими данными. В рамках курса вы узнаете, как с помощью современных методов и библиотек для анализа данных (OCR (компьютерное зрение), NLP (анализ языка), стек библиотек по работе с BigData и др.) решать задачи бизнеса. Программа ориентирована на достижение бизнес-результата, поэтому с первых занятий будут рассмотрены типичные аналитические задачи из разных индустрий, и, прежде чем перейти к изучению современных методов их решения, важно будет научиться понимать бизнес-смысл задачи, экономический эффект и какие инструменты применять для решения задач в зависимости от их специфики. Каждое занятие сопровождается практическими примерами и заданиями.

Участники, успешно завершившие курс, смогут самостоятельно писать основные алгоритмы машинного обучения, включая подготовку данных, отбор и формирование признаков, выбор оптимальных библиотек и их применение, а также выбор оптимальных метрик качества с учетом бизнес-задачи. Программа рассчитана на слушателей, имеющих знания в области линейной алгебры и теории вероятности на уровне первого курса технического вуза.

По результатам обучения у участников будет возможность поучаствовать в реальном состязании на Kaggle.

1. **Цель программы:**

Формирование знаний и навыков по решению основных задач машинного обучения, включая подготовку данных, отбор и формирование признаков, и по выбору оптимальных библиотек и их применению, а также выбору оптимальных метрик качества с учетом бизнес-задачи.

1. **Планируемые результаты обучения:**
   1. **Знание (осведомленность в областях):**
      1. Знать, что такое машинное обучение;
      2. Знать особенности решения задач машинного обучения;
      3. Знать особенности обработки больших данных;
      4. Знать алгоритмы ID3, C 4.5, CART
      5. Знать библиотеки для научных вычислений SciPy;
      6. Знать библиотеки машинного обучения Scikit-learn;
      7. Знать основы математики в машинном обучении (статистические распределения, нормализация признаков, приведение к нормальному распределению, метод максимального правдоподобия, проверка статистических гипотез, методы оптимизации);
      8. Знать метрики качества алгоритмов машинного обучения;
      9. Знать случай несбалансированных классов и случай классификации на несколько классов;
      10. Знать технику логистической регрессии;
      11. Знать метод опорных векторов;
      12. Знать способ машинного обучения – обучение без учителя;
      13. Знать основы теории графов, классические алгоритмы на графах;
      14. Знать алгоритм PageRank и алгоритмы поиска связных компонент в графе / сильно связных компонент в графе;
      15. Знать принципы применения на практике случайных веб-графов;
      16. Знать алгоритмы Apriori и FP-growth;
      17. Знать инструмент SPMF;
      18. Знать алгоритм коллаборативной фильтрации;
      19. Знать подходы Item-Based и User-Based к задаче рекомендации;
      20. Знать библиотеки NLTK\*;
      21. Знать модели представления текстов: мешок слов, VSM, синтаксические деревья;
      22. Знать современные методы: word2vec, topicmodelling;
      23. Знать подход машинного обучения MapReduce;
      24. Знать метод стохастического градиентного спуска;
      25. Знать концепцию вычислений в памяти и устойчивых распределенных наборов данных (RDD);
      26. Знать инструменты ApacheSpark;
      27. Знать основы функционального программирования (map, filter, reduce, lambda-функции);
      28. Знать библиотеки MLlib и GraphX Apache Spark;
      29. Знать альтернативы большим данным.
      30. Знать принципы построения нейронных сетей.
   2. **Умение (способность к деятельности):**
      1. Уметь решать задачи классификации, регрессии и кластеризации;
      2. Уметь осуществлять извлечение, отбор и преобразование признаков;
      3. Уметь работать с векторами и матрицами в библиотеке NumPy;
      4. Уметь визуализировать данные с Matplotlib и Seaborn;
      5. Уметь читать и обрабатывать данные с библиотекой Pandas;
      6. Уметь осуществлять первичный анализ данных с Seaborn;
      7. Уметь применять дерево решений Scikit-learn к синтетическому набору данных;
      8. Уметь решать задачи способом случайного леса, бустинга и бэггинга;
      9. Уметь осуществлять кластеризацию на графах/обнаружение в социальных сетях;
      10. Уметь применять машинное обучение в графовых задачах;
      11. Уметь обнаруживать знания в данных и анализировать последовательности;
      12. Уметь осуществлять поиск частых множеств, поиск сильно разделяющих паттернов, поиск паттернов с ограничениями, поиск частых последовательностей и поиск частых подграфов;
      13. Уметь оценивать качество рекомендательной системы;
      14. Уметь строить рекомендации на основе ассоциативных правил;
      15. Уметь использовать методы на основе матричной факторизации (SVD, PLSA, LDA, BMF);
      16. Уметь решать задачи обработки естественного языка (NLP);
      17. Уметь делать предобработку текстов: лемматизация, стемминг, синтаксический и морфологический анализ;
      18. Уметь осуществлять поиск ключевых слов, определять сходства документов, кластеризировать тексты и искать похожие слова;
      19. Уметь осуществлять коллаборативную фильтрацию с ApacheSpark;
      20. Уметь использовать функции map, filter, reduce и lambda-функции;
      21. Уметь использовать методы работы с устойчивыми распределенными наборами данных (RDD);
      22. Уметь анализировать веб-логи с ApacheSpark;
      23. Уметь решать задачи обработки текстов и обработки изображений и видео (ComputerVision);
      24. Уметь решать классические задачи DeepLearning.
   3. **Навыки (использование конкретных инструментов):**
      1. Владеть навыками настройки environment (Anaconda, виртуальная машина);
      2. Владеть навыками работы с pandas, numpy, matplotlib, seaborn;
      3. Владеть навыками решения задач классификации и регрессии, в том числе продвинутыми методами;
      4. Владеть навыками решения задач классификации;
      5. Владеть инструментами решения задач машинного обучения;
      6. Владеть навыками решения алгоритмических задач на графах;
      7. Владеть навыками решения задач в KaggleInclass;
      8. Владеть навыками разработки рекомендательных систем;
      9. Владеть навыками обработки текстов и знать библиотеку NLTK;
      10. Владеть навыками работы с инструментами vowpalwabbit;
      11. Владеть навыками работы с инструментами Theano, TensorFlow, Keras.
2. **Категории слушателей:**
   1. Образование: среднее профессиональное и/или высшее образование.
   2. Квалификация: рекомендовано для слушателей, имеющих знания в области линейной алгебры и теории вероятности на уровне первого курса технического вуза.
   3. Наличие опыта профессиональной деятельности: не требуется.
   4. Предварительное освоение иных дисциплин/курсов/модулей: не требуется.
3. **Учебный план программы «Практический анализ данных»**

В рабочем учебном плане отображается логическая последовательность освоения дисциплин, обеспечивающих формирование компетенций. Указывается общая трудоемкость дисциплин, а также их общая и аудиторная трудоемкость в часах.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **№**  **п/п** | **Наименование**  **модулей** | **Всего часов** | **Виды учебных занятий** | | | **Форма контроля** |
| **Лекции** | **Практические занятия** | **Сам. работа** |
| 1 | Модуль 1 «Основы машинного обучения» | 25 | 11 | 10 | 4 | **Практическая работа** |
| 2 | Модуль 2. «Дополнительные главы машинного обучения» | 23 | 10 | 10 | 3 | **Практическая работа** |
| 3 | Модуль 3. «Работа с большими данными и максимизация точности на примере соревнования по анализу данных» | 24 | 10 | 12 | 2 | **Практическая работа** |
|  | **ИТОГО** | **72** | **31** | **32** | **9** | **Практическая работа** |

1. **Календарный план-график реализации образовательной «Практический анализ данных»:**

Дата начала обучения: 1 ноября 2020 года.

Дата завершения обучения: 22 ноября 2020 года.

Периодичность набора групп – 1 раз в месяц.

|  |  |  |  |
| --- | --- | --- | --- |
| **№**  **п/п** | **Наименование**  **модулей** | **Трудоёмкость (ак. час)** | **Сроки обучения** |
|
| 1 | Модуль 1 «Основы машинного обучения» | 25 | 01.11.2020 – 06.11.2020 |
| 2 | Модуль 2. «Дополнительные главы машинного обучения» | 23 | 07.11.2020 – 15.11.2020 |
| 3 | Модуль 3. «Работа с большими данными и максимизация точности на примере соревнования по анализу данных» | 24 | 16.11.2020 – 22.11. 2020 |
|  | **ИТОГО** | **72** | **01.11.2020 – 22.11.2020** |

1. **Учебно-тематический план образовательной программы «Практический анализ данных»:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **№**  **п/п** | **Наименование**  **модулей/тем** | **Всего часов** | **Виды учебных занятий** | | | **Форма контроля** |
| **Лекции** | **Практические занятия** | **Сам. работа** |
| 1 | **Модуль 1. «Основы машинного обучения»** | 25 | 11 | 10 | 4 | Зачёт |
| 1.1 | Введение в машинное обучение. Основные типы задач и методы их решения | 2,5 | 2,5 | - | - |  |
| 1.2 | Библиотеки и инструменты для анализа данных. Математика в машинном обучении | 2,5 | 1,5 | 1 | - |  |
| 1.3 | Настройка environment (Anaconda, виртуальная машина). Практика с pandas, numpy, matplotlib, seaborn. | 2,5 | - | 2 | 0,5 |  |
| 1.4 | Алгоритм Обучение с учителем. Задачи классификации и регрессии. | 2,5 | 1,5 | 1 |  |  |
| 1.5 | Оценка качества алгоритмов машинного обучения. | 2,5 | 1,5 | 1 |  |  |
| 1.6 | Решение задач классификации с Kaggle. Альтернативные метрики качества алгоритмов классификации. | 2,5 | - | 2 | 0,5 |  |
| 1.7 | Продвинутые методы классификации и регрессии. Переобучение. Теория. | 2,5 | 2,5 | - | - |  |
| 1.8 | Продвинутые методы классификации и регрессии. Переобучение. Практика. | 2,5 | - | 1 | 1,5 |  |
| 1.9 | Обучение без учителя. | 2,5 | 1,5 | 1 |  |  |
| 1.10 | Решение задач классификации и регрессии с Kaggle. Борьба с переобучением. | 2,5 | - | 1 | 1,5 |  |
| 2 | **Модуль 2. «Дополнительные главы машинного обучения»** | 23 | 10 | 10 | 3 | Зачёт |
| 2.1 | Анализ социальных сетей. Теория. | 2,5 | 2,5 | - |  |  |
| 2.2 | Анализ социальных сетей. Практика. | 2,5 | - | 2 | 0,5 |  |
| 2.3 | Решение алгоритмических задач на графах. | 2,5 | - | 2 | 0,5 |  |
| 2.4 | Обнаружение знаний в данных. | 2,5 | 2,5 | - | - |  |
| 2.5 | Рекомендательные системы. | 2,5 | 2,5 | - | - |  |
| 2.6 | Разработка собственной рекомендательной системы кинофильмов. Решение проблемы холодного старта, разработка метрик качества алгоритмов рекомендации. | 2,5 | - | 1,5 | 1 |  |
| 2.7 | Обработка текстов. Теория. | 2,5 | 2,5 | - | - |  |
| 2.8 | Обработка текстов. Практика. | 2,5 | - | 2,5 | - |  |
| 2.9 | Подробный обзор библиотеки NLTK, решение задач с Kaggle. | 3 | - | 2 | 1 |  |
| 3 | **Модуль 3. «Работа с большими данными и максимизация точности на примере соревнования по анализу данных»** | 24 | 10 | 12 | 2 | Зачёт |
| 3.1 | Введение в анализ больших данных и масштабируемое машинное обучение. Теория. | 2,5 | 2,5 | - | - |  |
| 3.2 | Введение в анализ больших данных и масштабируемое машинное обучение. Практика. | 2,5 | - | 2 | 0,5 |  |
| 3.3 | Альтернатива большим данным. Large Scale Machine Learning. Обзор инструмента vowpal wabbit. Теория и практика. | 2,5 | 1 | 1,5 | - |  |
| 3.4 | Практика с Apache Spark, разбор альтернативных методов работы с большими данными. | 2,5 | - | 2 | 0,5 |  |
| 3.5 | Соревнования по анализу данных. Теория. | 2,5 | 2,5 | - | - |  |
| 3.6 | Соревнования по анализу данных. Практика. | 2,5 | - | 2 | 0,5 |  |
| 3.7 | Разбор нестандартных задач с Kaggle. | 3 | - | 2,5 | 0,5 |  |
| 3.8 | Deep Learning. | 3 | 2 | 1 | - |  |
| 3.9 | Продуктовая аналитика. | 3 | 2 | 1 | - |  |
|  | **ИТОГО** | **72** | **31** | **32** | **9** | **Зачёт** |

1. **Учебная (рабочая) программа повышения квалификации «Практический анализ данных»:**

**Модуль 1. «Основы машинного обучения» (25 ак. часов)**

**Тема 1.1. Введение в машинное обучение. Основные типы задач и методы их решения (2,5 ак. часа)**

Введение в машинное обучение, необходимые навыки;

Задачи классификации, регрессии и кластеризации;

Извлечение, отбор и преобразование признаков;

Открытие соревнования Kaggle Inclass;

Особенности решения задач машинного обучения;

Обзор инструментов для решения задач машинного обучения;

Обзор графовых задач;

Особенности обработки больших данных;

Деревья решений;

Энтропия, прирост информации и неопределенность Джинни;

Алгоритмы ID3, C 4.5, CART.

**Тема 1.2. Библиотеки и инструменты для анализа данных. Математика в машинном обучении (2,5 ак. часа)**

Работа с векторами и матрицами в библиотеке NumPy;

Обзор библиотеки для научных вычислений SciPy;

Математика в машинном обучении: статистические распределения, нормализация признаков, приведение к нормальному распределению, метод максимального правдоподобия, проверка статистических гипотез, методы оптимизации;

Тетрадки Jupyter (IPython) для презентации материала, содержащего код;

Визуализация данных с Matplotlib и Seaborn;

Чтение и обработка данных с библиотекой Pandas;

Практика на визуализацию и предварительный анализ данных соревнования Kaggle «Titanic: Machine Learning from Disaster» c помощью Pandas;

Первичный анализ данных с Seaborn;

Решение задачи соревнования Kaggle «Titanic: Machine Learning from Disaster» c помощью Pandas.

**Тема 1.3 Семинар. (2,5 ак. часа)**

Настройка environment (Anaconda, виртуальная машина). Практика с pandas, numpy, matplotlib, seaborn.

**Тема 1.4 Обучение с учителем. Задачи классификации и регрессии (2,5 ак. часа)**

Применение дерева решений Scikit-learn к синтетическому набору данных и к данным соревнования Kaggle Inclass по автострахованию;

Настройка параметров дерева, кросс-валидация;

Пример извлечения признака для набора данных соревнования «Titanic: Machine Learning from Disaster»;

Пример извлечения признака для набора данных соревнования Kaggle Inclass по автострахованию;

Практика на применение дерева решений и случайного леса к набору данных соревнования «Titanic: Machine Learning from Disaster».

**Тема 1.5 Оценка качества алгоритмов машинного обучения (2,5 ак. часа)**

Обзор библиотеки машинного обучения Scikit-learn;

Работа с признаками – отбор, преобразование, построение;

Метрики качества алгоритмов машинного обучения - доля (accuracy), точность (precision), полнота (recall), F-score, ROC-кривая, AUC;

Случай несбалансированных классов;

Случай классификации на несколько классов;

Логистическая регрессия;

Метод опорных векторов (Support Vector Machine), ядра;

Практика:

Сравнение разных методов при решении задачи Kaggle Inclass по предсказанию типа выплат по автостраховке;

Примеры решения задач классификации и регрессии — наборы данных UCI;

Решение задачи Kaggle «Greek Media Monitoring Multilabel Classification (WISE 2014)» Станиславом Семеновым.

**Тема 1.6 Семинар. (2,5 ак. часа)**

Решение задач классификации с Kaggle. Альтернативные метрики качества алгоритмов классификации.

**Тема 1.7. Продвинутые методы классификации и регрессии. Переобучение. Теория (2,5 ак. часа)**

Нейронные сети, алгоритм обратного распространения ошибки;

Построение ансамблей алгоритмов;

Случайный лес (Random Forest);

Бустинг (boosting) и бэггинг (bagging), Xgboost;

Стекинг;

Переобучение, кросс-валидация, регуляризация;

Пример регуляризации для логистической регрессии.

**Тема 1.8. Продвинутые методы классификации и регрессии. Переобучение. Практика (2,5 ак. часа)**

Сравнение случайного леса, бустинга и бэггинга на наборах данных репозитория UCI;

Случайный лес на примере набора данных Titanic;

Случайный лес на примере набора данных по автострахованию;

Практика использования библиотек Lasagne NN и Xgboost;

Разбор решения задачи Kaggle «Otto Group Product Classification Challenge» Станиславом Семеновым и Gilberto Titericz (1 место).

**Тема 1.9. Обучение без учителя (2,5 ак. часа)**

Введение в обучение без учителя;

Задача кластеризации — алгоритм k-means;

Иерархическая кластеризация;

Спектральная кластеризация;

Плотностные методы кластеризации;

Методы снижения размерности пространства признаков: кластеризация, метод главных компонент (PCA);

Поиск выбросов и аномалий в данных - статистический подход, одноклассовая машина опорных векторов;

Пример поиска аномалий в данных;

Решение задачи Kaggle «Driver Telematics Analysis» по определению профиля вождения водителя.

**Тема 1.10. Решение задач классификации и регрессии с Kaggle. Борьба с переобучением (2,5 ак. часа).**

**Модуль 2. «Дополнительные главы машинного обучения» (23 ак. часа)**

**Тема 2.1. Анализ социальных сетей. Теория (2,5 ак. часа)**

Введение в теорию графов;

Классические алгоритмы на графах;

Поиск в ширину и поиск в глубину;

Алгоритм PageRank;

Алгоритмы поиска связных компонент в графе / сильно связных компонент в графе;

Кластеризация на графах/обнаружение в социальных сетях.

**Тема 2.2. Анализ социальных сетей. Практика (2,5 ак. часа)**

Применение машинного обучения в графовых задачах;

Решение задачи рекомендации друзей в социальных сетях (Link Prediction);

Введение в случайные и веб-графы и как они помогают на практике в реальных задачах;

Обзор инструментов для работы с графами.

**Тема 2.3. Решение алгоритмических задач на графах. (2,5 ак. часа)**

**Тема 2.4 Обнаружение знаний в данных (2,5 ак. часа)**

Поиск частых множеств (товаров) и ассоциативные правила;

Алгоритмы Apriori и FP-growth;

Поиск частых, сильноразделяющих паттернов (frequent diverse patterns и emerging patterns);

Поиск паттернов с ограничениями;

Поиск частых последовательностей;

Поиск частых подграфов;

Знакомство с инструментом SPMF;

Анализ последовательностей на примере демографических данных.

**Тема 2.5. Рекомендательные системы (2,5 ак. часа)**

Введение в коллаборативную фильтрацию;

Item-Based и User-Based подходы к задаче рекомендации. Выбор меры сходства;

Оценка качества рекомендательной системы;

Рекомендации на основе ассоциативных правил;

Методы на основе матричной факторизации (SVD, PLSA, LDA, BMF);

Мультимодальная кластеризация и рекомендации в фолксономиях;

Case-study: рекомендация радиостанций.

**Тема 2.6.** **Разработка собственной рекомендательной системы кинофильмов (2,5 ак. часа)**

. Решение проблемы холодного старта, разработка метрик качества алгоритмов рекомендации.

**Тема 2.7. Обработка текстов. Теория (2,5 ак. часа)**

Задачи обработки естественного языка (NLP);

Предобработка текстов: лемматизация, стемминг, синтаксический и морфологический анализ;

Модели представления текстов: мешок слов, VSM, синтаксические деревья;

Современные методы: word2vec, topic modelling.

**Тема 2.8. Обработка текстов. Практика (2,5 ак. часа)**

Поиск ключевых слов;

Определение сходства документов;

Кластеризация текстов;

Поиск похожих слов.

**Тема 2.9. Подробный обзор библиотеки NLTK, решение задач с Kaggle (3 ак. часа)**

**Модуль 3. «Работа с большими данными и максимизация точности на примере соревнования по анализу данных» (24 ак. часа)**

**Тема 3.1. Введение в анализ больших данных и масштабируемое машинное обучение. Теория (2,5 ак. часа)**

Машинное обучение: подход MapReduce, онлайн-обучение;

Стохастический градиентный спуск;

Концепция вычислений в памяти и устойчивых распределенных наборов данных (RDD);

Обзор инструмента Apache Spark;

Введение в функциональное программирование (map, filter, reduce, lambda-функции);

Обзор библиотек MLlib и GraphXApacheSpark;

Коллаборативная фильтрация с Apache Spark.

**Тема 3.2. Введение в анализ больших данных и масштабируемое машинное обучение. Практика (2,5 ак. часа)**

Практика использования функций map, filter, reduce и lambda-функций;

Практика использования методов работы с устойчивыми распределенными наборами данных (RDD);

Анализ веб-логов с Apache Spark;

Пример построения рекомендательной системы фильмов с Apache Spark MLlib на данных MovieLens;

Пример решения задачи классификации со Spark MLlib;

Предсказание кликов пользователей с Apache Spark.

**Тема 3.3. Альтернатива большим данным. Large Scale Machine Learning. Обзор инструмента vowpal wabbit. Теория и практика (2,5 ак. часа)**

Теория:

Что делать, если мало оперативной памяти;

Введение в онлайн обучение;

Метрики качества - progressive loss;

Выбор функции потерь;

Hashing trick;

Подбор параметров регуляризации при он-лайн-обучении;

Дообучение алгоритмов

Практика:

Обзор инструмента vowpal wabbit;

Пример решения задачи в 2 строки с помощью vowpal wabbit;

Обзор нестандартных параметров vw.

**Тема 3.4.** **Практика с Apache Spark, разбор альтернативных методов работы с большими данными (2,5 ак. часа)**

**Тема 3.5. Соревнования по анализу данных. Теория (2,5 ак. часа)**

Обзор платформы Kaggle;

Особенности задач в соревнованиях по машинному обучению;

Отличия задач соревнований по анализу данных от реальных бизнес-задач.

**Тема 3.6. Соревнования по анализу данных. Практика (2,5 ак. часа)**

Выход в Top10% LeaderBoard на Kaggle в реальном времени

**Тема 3.7. Семинар (3 ак. часа)**

Разбор нестандартных задач с Kaggle.

**Тема 3.8. Deep Learning (3 ак. часа)**

Введение в нейронные сети;

Обзор возможностей нейронных сетей;

Задачи обработки текстов;

Задачи обработки изображений и видео (Computer Vision);

Инструменты (Theano, TensorFlow, Keras);

Разбор классических задач Deep Learning.

**Тема 3.9. Продуктовая аналитика (3 ак. часа)**

Тестирование гипотез в продуктовой аналитике;

Маркетинговые метрики качества алгоритмов машинного обучения;

Оценка экономического эффекта моделей;

Основные B2C метрики: CAC, LT, LTV, ARPU, ARPPU;

Модели машинного обучения в CRM.

**Описание практико-ориентированных заданий и кейсов:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Номер модуля** | **Наименование практического занятия** | **Описание** |
| **1** | **Модуль 1. «Основы машинного обучения»** | Практический семинар 1.1 | Настройка environment (Anaconda, виртуальная машина). Практика с pandas, numpy, matplotlib, seaborn |
| Практический семинар 1.2 | Решение задач классификации с Kaggle. Альтернативные метрики качества алгоритмов классификации |
| Практический семинар 1.3 | Решение задач классификации и регрессии с Kaggle. Борьба с переобучением |
| **2** | **Модуль 2. «Дополнительные главы машинного обучения»** | Практический работа 2.1 | Анализ социальных сетей. Практика |
| Практический семинар 2.2 | Решение алгоритмических задач на графах |
| Практический семинар 2.3 | Разработка собственной рекомендательной системы кинофильмов. Решение проблемы холодного старта, разработка метрик качества алгоритмов рекомендации |
| Практический работа 2.4 | Обработка текстов. Практика |
| Практический семинар 2.5 | Подробный обзор библиотеки NLTK, решение задач с Kaggle |
| **3** | **Модуль 3. «Работа с большими данными и максимизация точности на примере соревнования по анализу данных»** | Практический работа 3.1 | Введение в анализ больших данных и масштабируемое машинное обучение. Практика |
| Практический семинар 3.2 | Практика с Apache Spark, разбор альтернативных методов работы с большими данными |
| Практический семинар 3.3 | Разбор нестандартных задач с Kaggle |

**8.Оценочные материалы по образовательной программе**

**8.1. Вопросы тестирования по модулям**

|  |  |  |
| --- | --- | --- |
| **№** | **Вопросы входного тестирования** | **Вопросы итогового тестирования** |
|  | 1. Назовите виды задач линейного программирования.  2. Дайте определения понятиям события и вероятность?  3. Назовите основные элементы комбинаторики?  4. Какие основные теоремы в теории вероятности вы знаете?5. Умеете ли вы работать с матрицами и совершать операции над ними (линейные, умножение, транспонирование)6. Дайте определение понятию определитель?7. Имеете ли вы базовое представление о «Python»? Как осуществляется настройка среды: установка Anaconda8. Назовите основные типы объектов в Python.  9. Назовите основные понятия теории графов?  10. Какие существуют подходы к анализу текстов? | По результатам похождения программы необходимо будет выполнить итоговое задание (см. п. 8.3) |

**8.2.** О**писание показателей и критериев оценивания, шкалы оценивания.**

**Критерии оценивания тестирования**:

«Зачёт» - получают обучающиеся, справившиеся с тестом 65-100% (10 – 15 верных ответов);

«Незачёт» - соответствует работе, содержащей менее 65% правильных ответов (1- 9 верных ответов).

**Критерии оценивания выполнения практического задания:**

«Зачёт» - получают обучающиеся, успешно справившиеся с заданием: варианты оценки - “полностью справился с заданием”, “справился с заданием в основном, цель достигнута, но есть несущественные недочеты”;

«Незачёт» - ставится в случае невыполнения задания: варианты оценки - “не справился с заданием совсем”, “справился с заданием частично, но целевые результаты не достигнуты”.

**8.3 Примеры контрольных заданий по модулям или всей образовательной программе.**

**Практическое задание**. Построить модель прогноза дохода для датасета ниже. Также провести исследование данных и ответить на указанные ниже вопросы.

Задание выполнить на языке Python. Приложить тетрадку Python с подробными комментариями.

**Данные:**

https://archive.ics.uci.edu/ml/datasets/Adult

Age – возраст, количественный признак

Workclass – тип работодателя, количественный признак

fnlwgt – итоговый вес обьекта, количественный признак

Education – уровень образования, качественный признак

Education\_Num – количество лет обучения, количественный признак

Martial\_Status – семейное положение, категориальный признак

Occupation – профессия, категориальный признак

Relationship – тип семейных отношений, категориальный признак

Race – раса, категориальный признак

Sex – пол, качественный признак

Capital\_Gain – прирост капитала, количественный признак

Capital\_Loss – потери капитала, количественный признак

Hours\_per\_week – количество часов работы в неделю, количественный признак

Country – страна, категориальный признак

Вопросы:

1. Сколько мужчин и женщин (признак sex) представлено в этом наборе данных?

2. Каков средний возраст (признак age) женщин?

3. Какова доля граждан Германии (признак native-country)?

4. Постройте гистограмму распределения (bar plot) образования людей (признак education).

5. Каковы средние и среднеквадратичные отклонения возраста тех, кто получает более 50K в год (признак salary) и тех, кто получает менее 50K в год?

6. Правда ли, что люди, которые получают больше 50k, имеют минимум высшее образование? (признак education – Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters или Doctorate)

7. Выведите статистику возраста для каждой расы (признак race) и каждого пола. Используйте groupby и describe.

8. Среди кого больше доля зарабатывающих много (>50K): среди женатых или холостых мужчин (признак marital-status)?

9. Какое максимальное число часов человек работает в неделю (признак hours-per-week)? Сколько людей работают такое количество часов и каков среди них процент зарабатывающих много?

10. Посчитайте среднее время работы (hours-per-week) зарабатывающих много и мало (salary) для каждой страны (native-country), в частности, для Японии.

**8.5.** О**писание процедуры оценивания результатов обучения.**

Результаты выполнения практического задания оцениваются по достижению либо недостижению основной цели задания. Зачет- задание выполнено, цель достигнута в основном, н незачет- задание не выполнено, цель не достигнута.

Результаты ответов на вопросы оцениваются по проценту верных ответов: Зачет – более 65%, незачет – менее 65%.

**9.Организационно-педагогические условия реализации программы**

**9.1. Кадровое обеспечение программы**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **№**  **п/п** | **ФИО** | **Место основной работы и должность** | **Портфолио** | **Фото в формате jpeg** | **Отметка о полученном согласии на обработку персональных данных** |
| **1** | Марин Сергей | Преподаватель на курсе для менеджеров по особенностям применения «Big Data в бизнесе» | Опыт работы:  Основатель и глава департамента Больших Данных компании Вымпелком (бренд «Билайн»). Руководитель направления Business Intelligence и Data Mining ООО «Адидас», руководитель направления Business Intelligence в нидерландском KPN и ведущий менеджер продуктов в «Hewlett-Packard». Выпускник факультета ВМиК Московского Государственного Университета им. Ломоносова, выпускник программы MBA Московской Школы Управления Сколково |  | Получено |
| 2 | Семёнов Станислав | Преподаватель в специализации соревнования по анализу данных и машинное обучение, ранее первый в мире по рейтингу Kaggle | Консультант в области машинного обучения, номер 1 аналитик данных (Data Scientist) в мире по рейтингу Kaggle. Ранее HFT-аналитик в Quantstellation, преподаватель в Школе Анализа Данных Яндекс, портфельный аналитик в Тинькофф банк. Закончил Московский Физико-Технический Институт (МФТИ) и Высшую Школу Экономики (ВШЭ). Выпускник Школы Анализа Данных Яндекс |  | Получено |
| 3 | Игнатов Дмитрий | Преподаватель, специализация рекомендательные системы и Data Mining | Кандидат Наук в области прикладной математики и анализа данных от Российской Академии Наук и Технического Университета Дрездена, заместитель руководителя департамента анализа данных и искусственного интеллекта Национального Исследовательского Университета Высшей Школы Экономики (НИУ ВШЭ) |  | Получено |
| 3 | Сапунов Григорий | Преподаватель, специализация Глубокое Обучение, распознавание изображений и нейронные сети | Со-основатель и CTO в Intent.to, компании, специализирующейся на Глубоком Обучении и когнитивных сервисах. Ранее, преподаватель Когнитивных наук в Высшей Школе Экономики, руководитель направления в Яндекс. Выпускник Московской Школы Экономики |  | Получено |
| 3 | Алексеев Антон | Преподаватель, специализация анализ естественного языка и нейронные сети | Исследователь в совместном центре Искусственного Интеллекта Самсунг-Математического Института В.А.Стеклова. Ранее, разработчик библиотек работы с данными, систем анализа естественного языка и чат-ботов |  | Получено |

**9.2. Учебно-методическое обеспечение и информационное сопровождение**

|  |  |
| --- | --- |
| **Учебно-методические материалы** | |
| Методы, формы и технологии | Методические разработки,  материалы курса, учебная литература |
| <https://cocalc.com/>  <https://piazza.com/>  <https://www.anaconda.com/> | 1. Анализ больших наборов данных / Юре Лесковец, Ананди Раджараман, Джеффри Д. Ульман; пер. с англ. А.А. Слинкина. – Москва: ДМК Пресс, 2016. – 498 с. – Режим доступа: <http://znanium.com/catalog/product/1027845>  2. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / Виктор Майер-Шенбергер, Кеннет Кукьер (пер. с англ. Инны Гайдюк) изд. М.: Манн, Иванов и Фербер, 2014.  3. Убийственные большие данные / Кэти О’Нил (пер. В. Дягтерева) изд.: АСТ, 2017.  4. Теоретический минимум по Big Data / Анналин Ын, Кеннет Су (пер. А.В. Тимохин) изд.: Питер, 2019.  5. BIG DATA. Вся технология в одной книге. Андреас Вайгенд, 2018.  6. Основы Data Science и Big Data / Дэви Силен, Арно Мейсман, Мохамед Али, изд. Питер, 2018.  7. Укрощение больших данных / Билл Фрэнкс (пер. А. Баранов), изд.: Манн, Иванов и Фербер, 2014.  Big data простым языком / Алексей Благирев, Наталья Хапаева, изд. АСТ, 2019. |

|  |  |
| --- | --- |
| **Информационное сопровождение** | |
| **Электронные**  **образовательные ресурсы** | **Электронные**  **информационные ресурсы** |
| https://www.kaggle.com | https://dataschool.digital/blog |

**9.3. Материально-технические условия реализации программы**

|  |  |
| --- | --- |
| **Вид занятий** | **Наименование оборудования,**  **программного обеспечения** |
| Лекции и практические занятия | Севрвис видеоконференцсвязи с обменом сообщениями и передачей контента в режиме реального времени, с возможностью записи «Zoom» |
| Тестирования | Сервис «Google формы», облачные инструменты |
| Практические занятия | <https://cocalc.com/>  <https://piazza.com/>  <https://www.anaconda.com/> |

**Приложение 1**

1. ПАСПОРТ КОМПЕТЕНЦИИ

ОСНОВНАЯ ПРОФЕССИОНАЛЬНАЯ ОБРАЗОВАТЕЛЬНАЯ ПРОГРАММА ДОПОЛНИТЕЛЬНОЙ ПРОФЕССИОНАЛЬНОЙ ПОДГОТОВКИ

(повышение квалификации)

по направлению подготовки

**«Практический анализ данных**

ООО «ЮНИУМ»

УТВЕРЖДАЮ

Генеральный директор

ООО «Юниум»

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Д.С. Кондратьев

«8» октября 2020 г

|  |  |  |
| --- | --- | --- |
| **1.** | **Наименование компетенции** | **Умение выполнять анализ больших данных** |
| 2. | Указание типа компетенции | Профессиональная |
| 3. | Определение, содержание и основные сущностные характеристики компетенции | Формирование знаний и навыков по решению основных задач машинного обучения, включая подготовку данных, отбор и формирование признаков, а также по выбору оптимальных параметров алгоритмов и их применению, выбору оптимальных метрик качества и постановке задачи с учетом бизнес-задачи. |
| **4.** | **Дескриптор знаний, умений и навыков по уровням** | **Индикаторы** |
| Начальный уровень  (Компетенция недостаточно развита. Частично проявляет навыки, входящие в состав компетенции. Пытается, стремится проявлять нужные навыки, понимает их необходимость, но у него не всегда получается.) | Знать основы машинного обучения;  Знать особенности решения задач машинного обучения;  Знать основы графовых задач;  Знать особенности обработки больших данных;  Знать алгоритмы ID3, C 4.5, CART  Знать основы математики в машинном обучении, а также основы математического анализа, такие как статистические распределения, нормализация признаков, приведение к нормальному распределению, метод максимального правдоподобия, проверка статистических гипотез, методы оптимизации;  Знать, принципы организации нейронных сетей.  Уметь решать задачи классификации, регрессии и кластеризации;  Уметь осуществлять извлечение, отбор и преобразование признаков;  Уметь работать с векторами и матрицами в библиотеке NumPy;  Уметь визуализировать данные с Matplotlib и Seaborn;  Уметь читать и обрабатывать данные с библиотекой Pandas;  Уметь осуществлять первичный анализ данных с Seaborn;  Владеть навыками настройки environment (Anaconda, виртуальная машина).  Владеть навыками работы с pandas, numpy, matplotlib, seaborn;  Владеть навыками решения алгоритмических задач на графах. |
| Базовый уровень  (Уверенно владеет навыками, способен, проявлять соответствующие навыки в ситуациях с элементами неопределённости, сложности.) | Знать библиотеки для научных вычислений SciPy;  Знать библиотеки машинного обучения Scikit-learn;  Знать метрики качества алгоритмов машинного обучения;  Знать случай несбалансированных классов и случай классификации на несколько классов;  Знать сущность понятия логистическая регрессия;  Знать метод опорных векторов;  Знать, такой способ машинного обучения, как обучение без учителя;  Знать, основы теории графов, и знать классические алгоритмы на графах;  Знать альтернативы большим данным.  Уметь применять дерево решений Scikit-learn к синтетическому набору данных  Уметь решать задачи способом случайного леса, бустинга и бэггинга;  Уметь осуществлять кластеризацию на графах/обнаружение в социальных сетях.  Владеть навыками решения задач классификации и регрессии, в том числе продвинутыми методами;  Владеть навыками решения задач классификации;  Владеть инструментами решения задач машинного обучения. |
| Продвинутый  (Владеет сложными навыками, способен активно влиять на происходящее, проявлять соответствующие навыки в ситуациях повышенной сложности.) | Знать алгоритм PageRank и алгоритмы поиска связных компонент в графе / сильно связных компонент в графе;  Знать, что такое случайные веб-графы и как они применяются на практике;  Знать алгоритмы Apriori и FP-growth;  Знать инструмент SPMF;  Знать, что такое коллаборативная фильтрация;  Знать, что из себя представляют подходы Item-Based и User-Based подходы к задаче рекомендации  Знать библиотеки NLTK.  Знать модели представления текстов: мешок слов, VSM, синтаксические деревья.  Уметь обнаруживать знания в данных и анализировать последовательности;  Уметь осуществлять поиск частых множеств, частых, поиск сильно разделяющих паттернов, поиск паттернов с ограничениями, поиск частых последовательностей и поиск частых подграфов;  Уметь строить рекомендации на основе ассоциативных правил;  Уметь оценивать качество рекомендательной системы.  Уметь решать задачи обработки естественного языка (NLP);  Уметь делать предобработку текстов: лемматизация, стемминг, синтаксический и морфологический анализ;  Уметь осуществлять поиск ключевых слов, определять сходства документов, кластеризировать тексты и искать похожие слова.  Владеть навыками разработки рекомендательных систем;  Владеть навыками обработки текстов и знать библиотеку NLTK. |
| Профессиональный  (Владеет сложными навыками, создает новые решения для сложных проблем со многими взаимодействую-щими факторами, предлагает новые идеи и процессы, способен активно влиять на происходящее, проявлять соответствующие навыки  в ситуациях повышенной сложности.) | Знать современные методы: word2vec, topic modelling.  Знать подход MapReduce машинного обучения;  Знать стохастический градиентный спуск;  Знать концепцию вычислений в памяти и устойчивых распределенных наборов данных (RDD);  Знать инструменты Apache Spark;  Знать основы функционального программирования (map, filter, reduce, lambda-функции);  Знать библиотеки MLlib и GraphX Apache Spark.  Уметь использовать методы на основе матричной факторизации (SVD, PLSA, LDA, BMF);  Уметь осуществлять коллаборативная фильтрация с Apache Spark.  Уметь использовать функций map, filter, reduce и lambda-функций;  Уметь использовать методы работы с устойчивыми распределенными наборами данных (RDD);  Уметь анализировать веб-логи с Apache Spark;  Уметь решать задачи обработки текстов и обработки изображений и видео (Computer Vision)  Уметь решать классические задачи Deep Learning.  Владеть навыками решения задач по анализу данных в Kaggle Inclass;  Владеть навыками работы с инструментами vowpal wabbit.  Владеть навыками работы с инструментами Theano, TensorFlow, Keras. |
| 5. | Характеристика взаимосвязи данной компетенции с другими компетенциями/ необходимость владения другими компетенциями для формирования данной компетенции | Математика, теория вероятностей, программирование |
| 6. | Средства и технологии оценки | Практические задания  Тесты |

**VI. Рекомендаций к программе от работодателей**: наличие не менее двух писем и/или подтверждения на цифровой платформе Государственной системы предоставления ПЦС от работодателей о рекомендации образовательной программы для реализации в рамках Государственной системы предоставления ПЦС на формирование у трудоспособного населения компетенций цифровой экономики с указанием востребованности результатов освоения программы в сфере деятельности соответствующих компаний и готовности к рассмотрению заявок наиболее успешно освоивших образовательную программу граждан на прохождение стажировки и (или) собеседования на предмет трудоустройства путем проставления отметки в профиле программы **(Приложение 2).**

**V.Указание на возможные сценарии профессиональной траектории граждан** по итогам освоения образовательной программы (в соответствии с приложением):

По результатам прохождения образовательной программы возможны следующие сценарии профессиональной траектории развития граждан:

1. Лица, работающие по найму в организации, на предприятии или временно отсутствующие на рабочем месте смогу сохранить текущее место работы, развить свои профессиональные качества, повысить заработную плату и уровень дохода, сменить сферу работы без изменения сферы профессиональной деятельности, а также сохранить и развить свою квалификацию.
2. В рамках освоения данной программы у граждан появится возможность освоить новую востребованную профессию и устроиться на высокооплачиваемую работу, в том числе удалённо, зарегистрироваться в качестве самозанятых (фриланс), зарегистрировать ИП/открыть собственный бизнес.

**VIII.Приложенные Скан-копии утверждённой рабочей программы.**