

Программа учебной дисциплины
«Информационный менеджмент: Введение в Data Science»

Утверждена
Академическим советом ООП
Протокол № от «__»____20__ г.

Автор	Теванян Элен Арамовна, etevanian@hse.ru , Ульянкин Филипп Валерьевич, fulyankin@hse.ru Бабушкин Валерий Валерьевич, vbabushkin@hse.ru
Число кредитов	4 кредита
Контактная работа (час.)	46
Самостоятельная работа (час.)	106
Курс	1 курс бакалавриата
Формат изучения дисциплины	С использованием онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Цель учебной дисциплины познакомить студентов первого курса с культурой работы с данными, основными концепциями анализа данных и машинного обучения.

В результате освоения дисциплины студенты научатся:

- ставить измеримые цели;
- считать основные показатели;
- оценивать эффективность изменений;
- понимать, как делать прогнозы по данным.

Пререквизиты курса:

- знание английского языка на уровне Intermediate (B1/B2 по международной шкале)
- знание школьного курса алгебры и геометрии
- желательно знакомство с азами линейной алгебры, теории вероятностей и математической статистики

II. Содержание УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Введение в область Data Science.

Понятия Data Science, Machine Learning, Deep Learning, Big Data. Классы задач машинного обучения.

Тема 2. Математика для Data Science.

Линейная алгебра: векторы, матрицы. Теория вероятностей: вероятность, плотность, распределение, характеристики распределений. Математическая статистика: выборка, типы выборок.

Тема 3. Описательные статистики и визуализация данных.

Понятие описательных статистик. Минимум, максимум, среднее, стандартное отклонение, медиана, процентиля. Основные виды графиков.

Тема 4. Задача регрессия. Метрики регрессии. Линейная регрессия.
Постановка задачи регрессии. Метрики регрессии: MSE, MAE, MAPE, R²

Тема 5. Задача классификации. Метрики классификации.
Постановка задачи классификации. Метрики классификации: доля правильных ответов, точность, полнота.

Тема 6. Алгоритмы классификации: KNN, решающее дерево.
Алгоритм KNN. Алгоритм решающего дерева.

Тема 7. A/B-тестирование.
Понятие гипотезы, ошибок первого и второго рода. Тестирование гипотез.

Тема 8. Защита проекта с применением машинного обучения, или выход на инвестиционный комитет.

Тема 9. Кейсы машинного обучения в бизнесе: истории успехов и неудач.

III. ОЦЕНИВАНИЕ

Итоговая оценка за курс формируется нелинейно:

$$O_{\text{итог}} = \max\{0.7 \cdot O_{\text{накопленная}} + 0.3 \cdot O_{\text{Экзамен}}; 0.5 \cdot O_{\text{накопленная}} + 0.5 \cdot O_{\text{Экзамен}}\}$$

Накопленная оценка формируется из мероприятий текущего контроля:

$$O_{\text{накопленная}} = 0.1 \cdot O_{\text{DataCamp}} + 0.1 \cdot O_{\text{Семинары}} + 0.2 \cdot O_{\text{Самостоятельные}} + 0.2 \cdot O_{\text{Кейс}} + 0.2 \cdot O_{\text{Дз1}} + 0.2 \cdot O_{\text{Дз2}}$$

Таблица 1. Описание содержания мероприятий текущего контроля.

Оценка мероприятия текущего контроля	Содержание мероприятия текущего контроля
O _{DataCamp}	Оценка за изучение онлайн-курсов на платформе DataCamp. Шкала перевода представлена в таблице 2.
O _{Семинары}	Оценка за задачи, выданные на семинаре. Задач от 10 до 12 штук, каждая из которых оценивается по десятибалльной системе. Итоговая оценка по семинарам выставляется как среднее арифметическое по всем задачам.
O _{Самостоятельные}	Оценка за самостоятельные работы на семинарах. Планируется 4 работы, каждая из которых оценивается в десятибалльной шкале. Итоговая оценка по самостоятельным выставляется как среднее арифметическое по всем самостоятельным.
O _{Кейс}	Оценка за кейс. Кейс представляет собой групповую (по 3 студента) работу. Каждая группа получает оценку в десятибалльной шкале.
O _{Дз1}	Оценка за домашнее задание 1. Домашнее задание 1 представляет

	собой проектную работу группы студентов (по 3 студента), в котором нужно сделать ценовую сегментацию категории товаров. Каждая группа получает оценку в десятибалльной шкале.
Од32	Оценка за домашнее задание 2. Домашнее задание представляет собой индивидуальную работу по анализу трендов и проведению А/В-тестирования. Оценка выставляется в десятибалльной шкале.

Таблица 2. Шкала перевода оценок за DataCamp.

Количество выполненных заданий	Оценка в 10- балльной шкале
0 - 142	0
143 – 157	1
158 – 171	2
172 – 185	3
186 – 200	4
201 – 214	5
215 – 228	6
229 – 243	7
244 – 257	8
258 – 271	9
272 – 286	10

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

1. DataCamp

Название курса	Ссылка на курс
Introduction to Python For Data Science	https://www.datacamp.com/courses/intro-to-python-for-data-science
Intermediate Python For Data Science	https://www.datacamp.com/courses/intermediate-python-for-data-science
Pandas Foundations	https://www.datacamp.com/courses/pandas-foundations
Manipulating DataFrames with Pandas, Chapters: “Extracting and transforming Data”, “Grouping data”	https://www.datacamp.com/courses/manipulating-dataframes-with-pandas , chapters 16 4

Introduction to Data Visualization with Python	https://www.datacamp.com/courses/introduction-to-data-visualization-with-python
--	---

2. Пример семинарской задачи

- Попробуйте обучить метод одного ближайшего соседа. Что произошло с качеством модели? Как называется такая ситуация? Проинтерпретируйте её.
- Попробуйте перебрать соседей и узнать какое количество будет давать самое крутое значение ROC-AUC. Попробуйте сделать это с помощью цикла. Нарисуйте график, где по оси X будет отложено число соседей, а по оси Y значение ROC-AUC на тестовой выборке.

3. Пример задачи на самостоятельной работе

Найдите точность и полноту для предсказаний модели.

y_i	b_i
1	1
1	0
0	0
0	0

4. Пример кейса

В ваших руках оказались данные по удержанию сотрудников. Ваша задача состоит в предсказании того, уйдёт ли конкретный сотрудник с работы в ближайшее время. Одним из применений данной модели могла бы быть раздача ништяков тем людям, которые в ближайшее время хотят покинуть компанию. Возможно, вы могли бы попробовать понять какие именно ништяки нужно выписать сотрудникам из тестовой выборки, чтобы вероятность их оттока уменьшилась.

5. Домашние задания

Домашние задания 1 и 2 являются уникальными, детальные требования к их содержанию публикуются на вики-странице курса, ссылка на который представлена в разделе V.4.

6. Пример экзаменационного вопроса

Выберите правильный ответ на вопрос.

Какая из метрик является не подходит для задачи классификации?

- Точность
- Полнота
- Доля правильных ответов
- MSE

V. РЕСУРСЫ

1. Основная литература

1. David Julian, Designing Machine Learning Systems with Python, PACKT, 2016
2. Gene Kim, Kevin Behr, George Spafford, The Phoenix Project: A Novel About IT, DevOps, and Helping Your Business Win, IT Revolution Press, 2014
3. Jennifer Davis, Katherine Daniels, Effective DevOps: Building a Culture of Collaboration, Affinity, and Tooling at Scale, O'Reilly Media, Inc., 2016
4. Mark C. Layton, Agile Project Management For Dummies, John Wiley & Sons, 2012

2. Дополнительная литература

1. Как понять, что ваша предсказательная модель бесполезна;
<https://habrahabr.ru/post/337722/>
2. Метрики в задачах машинного обучения, <https://habrahabr.ru/company/ods/blog/328372/>
3. Байесовские многорукие бандиты против А/В тестов,
<https://habrahabr.ru/company/ods/blog/325416/>
4. crazyhatter. (2017). CRISP-DM: проверенная методология для Data Scientist-ов. Получено из Habrahabr: <https://habrahabr.ru/company/lanit/blog/328858/>
5. KDnuggets. (б.д.). Crisp-DM top methodology analytics. Получено из <http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>
6. Project Management Insitute, Inc. (2013). Руководство PMBOK 5.
7. Wikipedia. (б.д.). Cross Industry Standard Process for Data Mining. Получено из https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
8. Wikipedia. (б.д.). SEMMA. Получено из Wikipedia: <https://en.wikipedia.org/wiki/SEMMA>
9. НОУ "ИНТУИТ". (2017). Организационные и человеческие факторы в Data Mining. Получено из <http://www.intuit.ru/studies/courses/6/6/lecture/198>

3. Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Anaconda 2018.12 (Python 3)	Свободно распространяемое ПО

4. Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
<i>Интернет-ресурсы</i>		
1.	DataCamp	URL: https://datacamp.com
2.	Wiki-страница курса	URL: http://wiki.cs.hse.ru/Информационный_менеджмент:_Введение_в_Data_Science

5. Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

- ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);
- мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены компьютерами с установленной Anaconda (Python 3.6 и старше), с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.