Совместное применение Low- и High-code подходов в преподавании анализа данных

Введение

Курс по выбору “Предиктивная аналитика логистики и цепей поставок[1]” с 2017 года был включен в учебный план магистерской образовательной программы “Стратегическое управление логистикой и цепями поставок в цифровой экономике”. Цель дисциплины – формирование у студентов представления о том, как предиктивное моделирование используется в бизнесе, в сфере логистики и управления цепями поставок, а также формирование практических навыков использования инструментов и методов обработки данных и построения моделей. Первоначально предполагалось, что на курсе будут, в основном, обучаться выпускники бакалаврской программы “Логистика и управление цепями поставок”, которые уже достаточно хорошо владели бы инструментами анализа данных. В их учебном плане был годичный курс “Информационный менеджмент”, в котором они знакомились с базовыми инструментами и методами анализа (Excel, SPSS, Tableau, SQL), а также годичный научный семинар, на котором более углубленно изучался статистический анализ данных и прогнозирование в пакете R. Поэтому в рамках курса можно было бы поднять более сложные темы, такие как feature engineering, ансамбли моделей, глубокое обучение, фреймворк MLR для построения конвейеров машинного обучения. Однако на практике 30-50% слушателей курса, поступивших на программу с других ОП или из других вузов оказывались новичками в анализе данных. Не владея R, на старте курса они испытывали серьезные сложности с пониманием материала и выполнением заданий. Отчасти эту проблему удалось решить благодаря придания курсу статуса blended. Во время аудиторных занятий мы обсуждали методы предиктивного моделирования и бизнес-кейсы применения предиктивных моделей в логистике, а технические навыки обработки данных студенты получали в процессе изучения MOOC “Data Analysis with R”[2], кроме того, для курса был разработан репозиторий примеров на GitHub[3].

К сожалению, реформы образовательных программ по логистике в последние годы значительно ухудшили ситуацию с подготовкой студентов в области анализа данных – курсы по информационному менеджменту и семинар по анализу данных были исключены из программы, а количество часов на дисциплину для магистров сократилось на 25% (рис. 1). Отчасти ситуацию могла бы исправить подготовка студентов бакалавриата в рамках программы Data Culture, которая повсеместно внедрена в вышке и некоторых других университетах. Однако до того момента, пока ее слушатели дойдут до магистратуры, пройдет несколько лет, да и первые итерации курсов Data Culture, в которых еще не учтена специфика преподавания на разных ОП и не наработан опыт взаимодействия с их слушателями, могут давать не очень хорошие результаты в плане закрепления полученных навыков у студентов.

Рис. 1. Динамика количества аудиторных часов на дисциплину “Предиктивная аналитика логистики и цепей поставок”

[1] Программа курса https://www.hse.ru/edu/courses/480602487
[2] https://www.udacity.com/course/data-analysis-with-r--ud651
[3] https://github.com/postlogist/course_pan

В итоге, в 2020/21 учебном году, мы оказались в ситуации, когда одновременно сократилась как доля имеющих базовую подготовку студентов по анализу данных в R, так и количество часов на дисциплину. Потребовалось потратить много времени на объяснение базовых приемов обработки данных в R (трансформация/визуализация), «выпал» раздел про глубокое обучение, а по итогам курса было много негативных отзывов из-за чрезмерно сжатой программы. По этой причине в 2021/22 году был проведен редизайн курса, описанный далее.

Новая концепция курса и образовательные практики

В 2021/22 учебном году основную массу слушателей курса составили выпускники других образовательных программ, а объем был сокращен до 24 часов. Несмотря на это в рамках курса хотелось бы все же обсудить как современные методы анализа данных, так и практику их использования в бизнесе. Поскольку при использовании кода (R/Python) на начальном этапе неизбежно возникают сложности у не знакомых с такими инструментами студентов, было решено использовать в курсе на начальном этапе low-code подход. В качестве инструмента была выбрана система Orange Data Mining[4], специально разработанная для нужды обучения неподготовленных слушателей основам предиктивного моделирования. Эта система предлагает графический интерфейс для разработки сценариев обработки данных (рис. 2) и позволяет легко загружать данные, предобрабатывать и трансформировать их, выполнять разведочный анализ, строить и сравнивать различные модели. Orange может работать как со структурированными данными, так и с неструктурированными (тексты, изображения). Это инструмент с открытым исходным кодом, имеются дистрибутивы под Windows и MacOS. Она проще других аналогичных систем, например RapidMiner, KNime, но при этом достаточно функциональна.

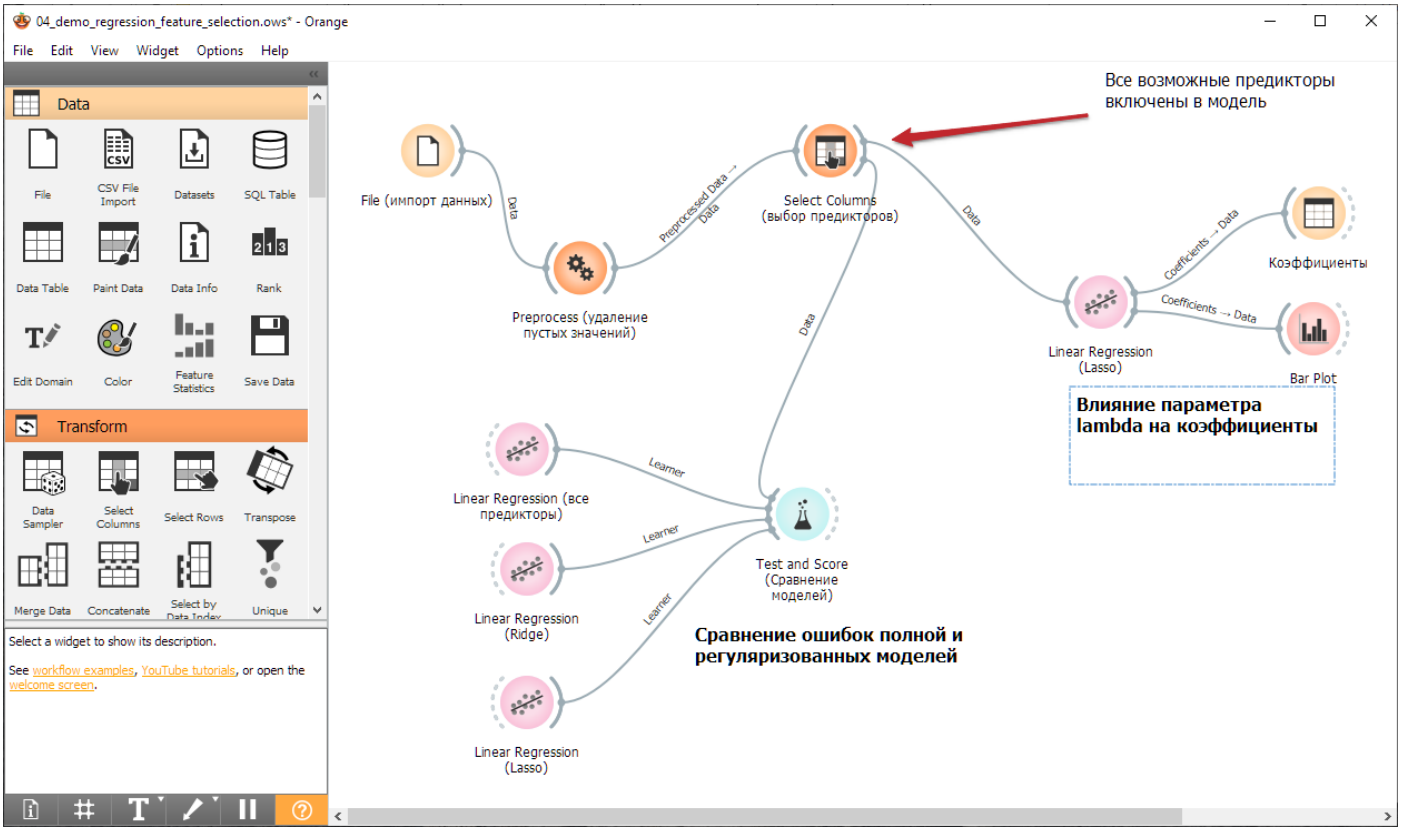

Рис. 2. Пример сценария в Orange для сравнения качества моделей и оценки влияния регуляризации

Благодаря low-code подходу на старте курса полностью исключаются проблемы с недостаточным владением инструментами у новичков, во время занятий можно в интерактивном режиме демонстрировать, как подготовка данных, различные спецификации моделей или параметры алгоритмов их обучения влияют на качество прогноза. В системе также хорошо реализовано сравнение моделей с помощью кросс-валидации.

[4] https://orangedatamining.com/

Работа с инструментом не вызывает сложностей у студентов, и фокус обсуждений смещается от проблем написания кода на непосредственно разведочный анализ, построение и интерпретацию моделей.

Конечно, разнообразие моделей и инструментов подготовки данных в Orange меньше, чем в R и Python. Также этот инструмент не предоставляет никаких средств для оптимизации параметров алгоритмов обучения моделей, кроме изменения их вручную. По этой причине в рамках курса студенты получают также базовые знания R и фреймворков для обработки данных (tidyverse и MLR), однако благодаря доступности Orange теперь они могут их изучать в гораздо более спокойном темпе. На рис. 3 показан календарный план курса по неделям. Параллельно с посещением аудиторных занятий студенты самостоятельно изучают рекомендованный MOOC. Тематически разделы курса согласованы с тематикой офлайн-занятий, например в первых работах студентам необходимы методы разведочного анализа и построения регрессионных моделей. На офлайн-занятиях мы решаем эти задачи в low-code инструменте, а в курсе студенты знакомятся с тем, как эти же задачи можно решить в R и привыкают писать код. Дополнительно к этому студенты могут изучать примеры решения задач курса на R в репозитории на GitHub[5].

На офлайн занятиях мы переходим к использованию R в самом конце, когда основные понятия, связанные с предиктивным моделированием им уже знакомы. Основанные на коде инструменты позиционируются как ценные средства, значительно расширяющие возможности аналитика.

Самостоятельная работа студентов организована с помощью домашних заданий. Два из них основаны на материалах офлайн-курса и предполагают обработку данных в Orange, два – на материалах MOOC.

Важной частью курса является обсуждение практики использования предиктивного моделирования в сфере логистики и управления цепями поставок. Студентам предлагаются материалы для подготовки докладов по бизнес-кейсам использования предиктивного моделирования[6]. На каждом занятии, начиная со второго, группа студентов может выступить с докладом, после которого проводится дискуссия с учебной группой. Эта практика позволяет повысить интерес студентов к изучению анализа данных, их вовлеченность на занятиях, а также убеждает их в том, что данное направление является актуальным трендом в бизнес-среде.

Для закрепления пройденного в рамках курса материала студентам предлагается реализовать групповой проект, в котором они могут либо решить свою задачу в рамках исследовательской или профессиональной деятельности, либо поучаствовать в соревновании по анализу данных (например, на платформе Kaggle, DrivenData или Crowd Analytix, AnalyticsVidhya). Защита проекта выносится на устный экзамен, чтобы обеспечить достаточное время на его реализацию.

В результате применения описанных практик в рамках курса были достигнуты следующие результаты:

- Доля сданных в срок заданий увеличилась

- Повысилось качество выполнения итоговых проектов по курсу, студенты демонстрируют владение более широким кругом методов обработки данных

-- Несмотря на значительное сокращение времени на изучение дисциплины, в СОП нет негативных комментариев о чрезмерно сжатой программе, а оценки СОП сохранились на том же уровне

- Курс выбран студентами как лучший по критериям "Полезность для расширения кругозора и разностороннего развития" и "Новизна полученных знаний".

«Гибридный» подход, при котором для обучения анализу данных используются как Low-code инструменты, так и код на R/Python может быть полезен прежде всего на курсах с очень разнородной по уровню подготовки аудиторией (например, на магистерских программах). В похожей ситуации я оказался, например, на курсе “Наука о данных для бизнеса” для ОП “Управление цифровым продуктом”, на котором есть как практикующие дата-аналитики, так и люди с базовым художественным образованием, никогда не имевшие

[5] https://github.com/postlogist/course_pan
[6] https://github.com/postlogist/course_pan/blob/master/presentation_topics.md

дело с данными. Важно и то, что часть материала переводится в blended-формат, поскольку слушателям с разной подготовкой необходим разный объем теоретического материала. Вариантов реализации этого довольно много – можно подобрать подходящий по контенту MOOC, собрать набор курсов в DataCamp Classrooms[7] или записать видеолекции самостоятельно[8].

Среда, тема/Неделя	**1**	**2**	**3**	**4**	**Каникулы**		**5**	**6**	**Работа над проектом**		**Экзамен**
Офлайн											
Виды данных. Концепция статистического обучения. Модели, оценка качества модели.											
Поиск зависимостей в данных. Разведочный анализ в Orange											
Задача регрессии. Построение и интерпретация моделей.											
Конструирование и отбор признаков для предиктивных моделей (практикум)											
Задача классификации. Метрики качества моделей классификации.											
Практикум по классификации. Ансамбли моделей, объяснимые модели машинного обучения.											
Конвейеры обработки данных в R: tidyverse, MLR											
Обсуждение кейсов предиктивного моделирования в логистике											
ДЗ 1. Построение модели регрессии (Orange)											
ДЗ 2. Построение модели классификации (Orange)											
Выполнение и защита проекта											
MOOC											
Понятие разведочного анализа, применение											
Основы R											
Разведочный анализ для 1 переменной											
Разведочный анализ для 2 переменных											
Разведочный анализ для множества переменных											
Построение предиктивных моделей в R											
ДЗ 3. Разведочный анализ для 1 и 2 переменных в R											
ДЗ 4. Построение и интерпретация моделей регрессии в R											

Рис 3. График обучения на курсе “Предиктивная аналитика логистики и цепей поставок”

[7] https://www.datacamp.com/groups/classrooms

[8] https://www.youtube.com/playlist?list=PLwCnsQacFoW5CnPVPDj2uSeFYcWhRe0v7 – плейлист с видео для моего курса Наука о данных для бизнеса, где используется Orange в качестве основного инструмента