Государственное бюджетное общеобразовательное учреждение

средняя общеобразовательная школа №138

имени Святого благоверного князя Александра Невского

Калининского района Санкт-Петербурга

Курсовая работа

«Наука о данных в сфере образования»

Выполнил: Ученик 9а класса

Фиолетов Эдуард

Руководитель:

О.О.О

Санкт-Петербург

2020

Содержание

1. Вступление

2. Основная часть

2.1 Теоретическая часть

2.1.1 Что такое наука о данных?

2.1.2 Как работает Аналитик данных

2.1.3 Составляющие Data Science

2.1.4 Инструменты в работе

2.2 Практическая часть

2.2.1 Постановка задачи

2.2.2 Собираем исходные данные

2.2.3 Предварительный анализ

2.2.4 Нахождение значимостей

2.2.5 Точка прогноза

2.2.6 Построение модели

2.2.7 Результат

3. Заключение

4. Список литературы

4.1 Ссылки

4.2 Литература

1. Вступление

*Отправной точкой в этом проекте стала 1 неделя карантина, когда мне попалась реклама "Яндекс Практикум". Они обещали профессиональную переподготовку за несколько месяцев, на должность "Data Scientist", из которых первые две недели бесплатно. К концу бесплатного периода мне так сильно понравилась эта область, что ко мне стало приходить понимание – это будет кормить не только меня, но и окружающих. Все лето я изучал дисциплины этой отрасли, и понял, что это задача на всю жизнь.*

Спорим не угадали автора этого вступления? Да всё правильно, его писал не я, а одна из самых продвинутых нейросетей по обработке естественного языка (NLP) – «GPT-3Large» под разработкой международной компании «OpenAI», с дообучением нашего «Сбера».

Да, к сожалению, мне не выдали разрешение на использования API от компании основателя, на авторы выложили исходный код в общий доступ, чем и пользуются различные компании и отдельные разработчики, модернизируя изначальную модель.

Вся моя работа заключалась в том, чтобы указать заголовок, внутренние теги, опорные точки текста (имена, даты, места), размер нужного материала, и дополнительные настройки (использование ненормативной лексики, длинна чисел с плавающей точкой etc.)

Даже если вы раскусили меня, и нашли, чем машина себя выдала, есть ещё множество областей, где человек уже бессилен. Например, разработчики подразделения компании «NVIDIA» - «NVlabs», создали генеративно-состязательную нейросеть, которая создаёт изображения людей, которых никогда не существовало. В большинстве случаев, вероятность угадывания работы нейросети среди реальных фотографий, равняется случайной. То-есть при отсутствии видимых артефактов, вы не сможете сказать наверняка, где реальный человек. Также существуют нейросети, генерирующие текста и музыку для музыкантов, картины в стиле современного искусства, самих деятелей культуры и ещё много чего ещё.

К чему я это всё? Да к тому, что по-факту, машина прошла «Тест Тьюринга»!

А что это значит? А значит это то, что будущее уже стучится в дверь, и просит, чтобы ты уже начал изучать науку о данных - Data Science

А конкретно в этой работе я буду рассматривать места применения машинного обучения в школах, вузах и других образовательных учреждениях.

2. Основная часть

2.1 Теоретическая часть

2.1.1 Что такое наука о данных?

“[Наука о данных](http://bit.ly/quaesita_datasci) — это дисциплина, позволяющая сделать данные полезными” - это мое любимое определение науки о данных.

Ещё есть цитата википедии:

“Раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме. Объединяет методы по обработке данных в условиях больших объёмов и высокого уровня параллелизма, статистические методы, методы интеллектуального анализа данных и приложения искусственного интеллекта для работы с данными, а также методы проектирования и разработки баз данных.”

Также, чтобы было понятней чем же занимается специалист по анализу данных, можно привести несколько примеров задач, которые ставят перед аналитиком:

1. Обнаружение аномалий, например, ненормальное поведение клиента, мошенничества

2. Персонализированный маркетинг — электронные рассылки, ретаргетинг, системы рекомендаций

3. Количественные прогнозы — показатели эффективности, качество рекламных кампаний и других мероприятий

4. Скоринговые системы — обработка больших объёмов данных, помощь в принятии решений, например, о предоставлении кредита

2.1.2 Как работает Аналитик данных

Уже из определения вы могли понять, что тут очень много ***различной*** работы. Поэтому следует прояснить основные чести работы data scintist’а:

1. Формулировка вопроса

Оценка масштабов, времени, сложности

2. Предобработка данных

Очистка сырых данных, анализ достоверности данных, research

3. Анализ чистых данных

Самая интересная часть работы. Разрабатываем модели.

4. Визуализация

Представление итоговой информации в красивых графиках

5. Реакция

Принятие решения на основе итогов анализа

В реальном мире очень мало специалистов, которые могут эффективно работать во всех областях, поэтому для каждой дисциплины существуют более узкие специальности:

* Найти проблему - Аналитик BI
* Поставить задачу - Data Analyst
* Собрать данные - Data Engineer
* Провести аналитику - Data Science
* Разработать предиктивную модель - ML Engineer

2.1.3 Составляющие Data Science:

1. Математика

В Data Science из царицы наук берётся 4 раздела: Статистика, Линейная алгебра, Теория вероятностей, Мат. Анализ. Эти фундаментальные математические знания действительно важны, чтобы анализировать результаты применения алгоритмов обработки данных. Сильные инженеры в машинном обучении без такого образования есть, но это скорее исключение.

2. Программирование

По факту, в этой сфере абсолютную победу одержал язык программирования “Python”. Почему? Ответ прост: Потому-что “питон” - самый простой. Но в промышленные модели машинного обучения переписываются на более быстрых языках – например, C++ или Golang.

3. Data Mining

Важный исследовательский процесс. Он включает анализ скрытых моделей данных в соответствии с различными вариантами перевода в полезную информацию, которая собирается и формируется в хранилищах данных для облегчения принятия деловых решений, призванных сократить расходы и увеличить доход

2.1.4 Инструменты в работе

* Базы данных, таблицы, и другие формы хранения информации (SQL, CSV, EXCEL, NOSQL, etc)
* Математические алгоритмы (Линейная регрессия, Дерево решений, Логистическая регрессия, Random forest, Градиентный бустинг, etc)
* Программирование профессиональных библиотек (scikit-learn, pandas, numpy, tensorflow, scipy, keras, etc)
* Построение красивых графиков (dashboards, matplotlib, seaborn, plotnine, plotly)
* Soft-skills чтобы донести результаты своей работы до руководства

2.2 Практическая часть

2.2.1 Постановка задачи

Сейчас Google способен определять развитие эпидемий по поисковым запросам, сервис онлайн знакомств Eharmony по анкетам пользователей определяем вероятность создания семьи, банки с высокой точностью определяют риск невозврата кредита по документам заявителя.

Но возможно ли при помощи данных предсказать исход обучения студентов Вуза? Получит конкретный студент диплом или будет отчислен?

Итого наш вопрос: «Можно ли предугадать успешность обучения студента?»

2.2.2 Собираем исходные данные

Так как я сам учусь еще в школе, просить собрать исходные данные мне пришлось у своих друзей из Санкт-Петербургского Политехнического университета. Спустя несколько месяцев неспешной работы по выходным, мне прислали данные процесса обучения нескольких десятков выпускников этого Вуза.

Итак, в нашем распоряжении аккуратные обезличенные данные о исходе обучения, поле, возрасте, кафедре, номере группы, оценках за сессию, форме оплаты, отметках о продлении сессии, наличии и размеру стипендии.

2.2.3 Предварительный анализ

Первичный анализ позволяет взглянуть на проблему общим взглядом и выработать первоначальные гипотезы. Вот, что удалось выяснить:

* Треть от числа поступивших на первый курс отчислят в течение обучения
* Большинство студентов отчисляют на I курсе
* Среди отчисленных больше мальчиков
* Вероятность закончить ВУЗ очень сильно отличается на разных кафедрах (от 0.1 до 0.65)

2.2.4 Нахождение статистических значимостей

Этот шаг - самый важный в анализе данных. Аналитику нужно отобрать признаки, которые несут в себе сигналы об отчислении и выбрать модель, которая сможет продемонстрировать лучший результат, учитывая специфику данных (размер выборки, тип переменных и многое другое). После проведения статистических тестов выявлено, что о скором отчислении могут сообщить следующие признаки: средний балл на сессии, наличие стипендии, форма оплаты, кафедра и пол.

2.2.5 Точка прогнозирования

Выбор точки прогнозирования подразумевает выбор того момента в процессе обучения студента, когда накопленных про него данных будет достаточно для осуществления точного прогноза. Очевидно, что прогнозирование отчисления студентов после окончания 3 курса не имеет практической ценности, а данных, полученных еще до начала обучения недостаточно для построения моделей.

Оказалось, что модели построенные на данных после первого курса уступают в точности моделям, построенным на данных после 2 курса лишь несколько процентов, но очевидно имеют гораздо большую практическую пользу.

2.2.6 Построение модели

Построение прогностических моделей происходит следующим образом. Имеющиеся данные о студентах и исходах их обучения разбиваются на две части: обучающую и тестовую выборку. Первая нужна для того, чтобы натренировать алгоритм. Современные математические модели способны чутко улавливать сочетания показателей, которые свидетельствуют от скором отчислении студентов.

Например, получить диплом "Кафедры Пик" сама по себе сложная задача. Для молодых людей, которые учатся на коммерческой основе и имеют долю прогулов выше определенного порога - это практически нереально.

После обучения алгоритма точность его прогнозов оценивается на тестовой выборке (часть исходных данных, которая не использовалась для "обучения").

Алгоритм, который демонстрирует самую низкую ошибку по целому спектру индикаторов и выбирается в качестве рабочего. Часто самое точно решение дает не отдельная модель, а усредненный прогноз нескольких алгоритмов (это называется бэггинг). Так и получилось в нашем случае. На тестовых выборках ансамбль из нескольких моделей верно предсказывает исход обучения 88% студентов.

2.2.7 Результат

Итак, итоговая модель (точнее ансамбль моделей) в 88% случаев верно определяет исход обучения студентов по данным, которые имеются о них на момент окончания I курса. Фактическую точность прогнозов можно будет проверить в следующем году, когда свои дипломы получат (или не получат) студенты из контрольной группы. На данный момент модель пророчит отчисление для трети из них.

Пока в Вузе нет единой информационной системы, процесс прогнозирования нельзя автоматизировать, а жаль. Ведь модель могла бы дообучаться на основе вновь поступающих данных.

Текущее решение может быть использовано в ручном режиме, при котором завуч готовит таблицу о студентах и их результатах после первого курса и загружает ее в программу. Рекомендации программы относительно вероятности отчисления можно использовать для проведения бесед со студентами из нижней части списка.

3 Заключение

Пока Вузы не хранят школьные анкеты абитуриентов, поэтому данных для предсказания исхода обучения на этапе вступительных экзаменов слишком мало. Но грядет время, когда с высокой точностью можно будет оценить вероятность получения диплома Вуза еще по школьным оценкам будущих студентов в сочетании с другими фактами их биографии...

4. Список литературы

Ссылки

https://openai.com/ - *сайт создателей нейросети “GPT-3Large”*

https://thispersondoesnotexist.com/ - *проект нейросети от “NVlabs” по созданию изображений людей, которых никогда не существовало*

https://ru.wikipedia.org/wiki/Наука\_о\_данных

https://netology.ru/blog/01-2020-gid-po-data-science

https://academy.yandex.ru/posts/chem-zanimaetsya-spetsialist-po-data-science-i-kak-nachat-rabotat-v-etoy-oblasti

https://neerc.ifmo.ru/wiki/

https://habr.com/ru/post/538562/

https://sbercloud.ru/ru/warp/gpt-3

https://www.tidydata.ru/predictive

Литература

Работа с данными в любой сфере: Как выйти на новый уровень, используя аналитику / Кирилл Еременко ; Пер. с англ. — М. : Альпина Паблишер,

2019 — 303 с. - ISBN 978-5-9614-2582-6

Python и машинное обучение: машинное и глубокое обучение с использованием Python, scikit-learn и TensorFlow 2, / Sebastian Raschka Vahid Mirjalili / 3-е изд.: Пер. с англ. - СПб. : ООО "Диалектика", 2020. - 848 с.: ISBN 978-5-907203-57-0