Applied Network Science, 2(1). -2017. - 15 p.
18. The Open Graph Viz Platform (Release 0.9.2/ 26 September 2017) [Электронный документ] URL: https://github.com/gephi/gephi (Дата обращения: 09.09.2019)

ИСПОЛЬЗОВАНИЕ ОБЛАЧНЫХ ТЕХНОЛОГИЙ МАШИННОГО ОБУЧЕНИЯ В СОЦИАЛЬНЫХ ИССЛЕДОВАНИЯХ С ОГРАНИЧЕННЫМ ФИНАНСИРОВАНИЕМ

С.В. Романчуков
(г. Томск, Томский политехнический университет)
e-mail: inoytomsk@yandex.ru

CLOUD-BASED MACHINE LEARNING PLATFORMS FOR SOCIAL RESEARCH LOW ON RESOURCES

S.V. Romanchukov
(Tomsk, Tomsk Polytechnic University)

Abstract: Actual paper compares machine learning cloud platforms that can be used by non-programmers, it is addressed to sociologists, psychologists, gender researchers. Their work can require more of a data scientific approach, than their funds can allow to support. This tools can be used with minor expertise in the field of machine learning, programming and mathematics and don't require professional data scientists and analysts in team. We take one testing data set similar to those generated in field-level social studies and fallow through every stage of model training and tweaking in Google Cloud AI and IBM Watson Auto AI - two cloud0based platforms for data science without engineering.

Keywords: cloud platform, classification, machine learning

Введение. Машинное обучение и искусственные нейронные сети достаточно эффективны в решении задач классификации, распознавания образов, прогнозирования поведения сложных систем и выбора неизвестных параметров, которые связаны с характеристиками сложных объектов, в том числе социально-экономических систем [1]. Существует несколько методов так называемого «обучения» нейронной сети: обучение с учителем, обучение без учителя, обучение с подкреплением [2].

Первый вариант - обучение с учителем (способ реализации машинного обучения, при котором тестовая система подвергается принудительному обучению с использованием примеров «стимул-реакция») - кажется наиболее подходящим для решения проблемы взаимосвязи между двумя группами параметров, классификация или прогнозирования параметров[3].

Метод обучения с учителем предполагает, что может быть некоторая неустановленная связь между входными и выходными данными. Известен только конечный набор прецедентов - пар стимул-реакция - называемая обучающим набором. На основе этих данных происходит итеративный процесс выбора параметров с целью восстановления зависимости и построения подходящей для прогнозирования модели отношений [4].

Организация процесса обучения с нуля является нетривиальной математической и программистской задачей, однако на данный момент на рынке доступно значительное количество облачных программных продуктов, которые позволяют создавать проекты AI и ML без специальных знаний. Исследовательские группы с ограниченным финансированием, лишённые доступа к профессиональным знаниям в этой области могут использовать такие решения, однако, чтобы подходить упомянутой аудитории, программный продукт должен отвечать нескольким дополнительным требованиям, кроме точности и технических параметров:

- Простая и доступная документация и пользовательский интерфейс
- Простота использования и интерпретации результатов

- Доступные цены, возможности и функционал пробных версий

В качестве исходного массива данных, с которым сравнивались системы, была взята выборка компаний, включающая данные за период с 1 января 2018 года по сентябрь 2019 года - 1640 строк данных и 24 переменных с описанием их коммерческих показателей, параметров их рекламных кампаний и внутренних социальных связей, показатели, относящиеся к трудовой дисциплини, включая, но не ограничиваясь:

- Название и доменный адрес компании
- Индустрия и
- Годовая выручка
- Основной рекламный канал
- Затраты на рекламу (аккумулированные из источников SEMRush и SPYfu по-отдельности)
- Количество рекламных креативов
- Ключевые рекламные пиксели
- Количество сотрудников
- Динамика численности личного состава
- Оценки атмосферы в коллективе и условий труда

Этот набор числовых и категориальных переменных, отражающих как экономическую, так и социальную политику внутри компании, использовался для прогнозирования одной логической переменной - выполняет ли эта компания одно конкретное действие или нет. Это классическая задача бинарной классификации - она может быть решена с помощью алгоритмов машинного обучения и может служить хорошей областью тестирования для сравнения различных платформ.

Концепция MLaaS. Концепция машинного обучения как услуги (MLaaS) уже используется для маркетинговых исследований и работы с большими данными в коммерческих структурах, но все же до сих пор недооценивается в научных организациях. MLaaS - это общее определение для различных облачных платформ, которые осуществляют предварительную обработку данных, обучение модели и оценку модели для дальнейшего прогнозирования. Существует целый ряд платформ MLaaS, которые обеспечивают быстрое обучение и развертывание моделей, из которого выделяются четыре ведущих облачных сервиса MLaaS [5]:

- Google Cloud AI
- IBM Watson Auto AI
- Amazon Machine Learning
- Azure Machine Learning

В этой статье мы сконцентрируемся на первых двух платформах, исходя из предположения, что продукты Google и IBM ближе к программному обеспечению, уже широко распространенному в социальных исследованиях (например, Google Forms, Google Sheets, IBM SPSS), и совместимы с данными, обрабатывавшимися ранее с помощью этого программного обеспечения. Обе платформы реализуют схожую последовательность операций (т.н. "стандартный воркфлоу машинного обучения") [6][7]:

- Подготовка данных;
- Обучение модели;
- Оценка качества модели по предоставленному набору метрик;
- Тестирование модели;
- Развёртывание и начало эксплуатации.

Обе системы позволяют решать сходный набор задач в области:

- Бинарной классификации,
- Мультиклассовой классификации,
- Построения регрессионной модели.

Watson предоставляет немного больше возможностей для выбора алгоритмов решения и их параметров, но Google обладает более удобным пользовательским интерфейсом - с очень подробной встроенной справкой. Для работы в Auto AI требуется меньше опыта - почти у любого элемента, термина или диаграммы рядом есть подсказка, дающая краткое, но информативное пояснение. В Auto ML аналогичные компоненты присутствуют без пояснений.

Ценовая политика платформ. Обе платформы основывают свои цены на количестве времени и количестве узлов, необходимых для выполнения задачи.

Google AutoML Tables предоставляет бесплатную пробную версию, в 6 бесплатных часов обучения и прогнозирования для каждой учетной записи. После этого цены зависят от типа операции:

- Обучение модели - $ 19,32 за час вычислительных ресурсов
- Развертывание модели - $ 0,005 за ГБ/час/ машину (модель дублируется на 9 машин)
- Пакетное прогнозирование стоит $ 1,16 за час
- Онлайн-прогнозы стоят $ 0,21 за час

Модель ценообразования IBM Watson Auto AI основана на так называемом подходе к единице емкости (CUH), который позволяет подбирать различные типы оборудования. Например, 4 часа использования однойNVIDIA K80, час NVIDIA V100 или час автоматического прогнозирования на машине 16 vCPU и 64 ГБ ОЗУ стоит равное количество CUH.

Минимальный тарифный план IBM Watson Machine Learning полностью бесплатный и предоставляет максимум пять развернутых моделей, 5000 прогнозов в месяц и 50 единиц емкости в месяц. Стандартный предполагает плату в размере - 0,54 доллара США за один CUH или 1000 прогнозов.

В ходе общего тестирования обоих продуктов в одном и том же наборе данных было проведено 1,81 ч обучения модели на платформе Google (30% от общего объема пробных ресурсов или ~ $35 платного), на платформе Watson те же тесты заняли 26% бесплатного месячного лимита.

Подготовка данных. Обе среды позволяют импорт данных:

- Из собственной инфраструктуры данных (Google Cloud / IBM Cloud)
- Из различных интегрированных источников данных и баз данных
- С персонального компьютера в виде отдельного файла

Чтобы сымитировать наиболее вероятный для небольшой исследовательской группы или студенческого проекта вариант, мы загружали данные с ПК в формате .CVS - одном из самых популярных форматов файлов для большинства аналитических программ на основе таблиц [8].

Обе платформы поддерживают CSV-импорт и по умолчанию предлагают автоматическое распознавание типов переменных. Обе платформы загрузили данные достаточно быстро, но продукт Google обрабатывал данные почти в 5 раз медленнее (2,5 минуты против менее чем 30 секунд для Watson). С другой стороны, IBM Watson не смог распознать две категориальные переменные с большим количеством категорий и пометил их как уникальные текстовые строки.

Интегрированный компонент для ручной подготовки и улучшения данных доступен для обеих этих платформ, так что это не является серьезной проблемой, но требует ручного вмешательства.

Обучение и оценка моделей. После загрузки данных и выбора цели обучения оба решения предоставляют набор дополнительных опций. Google больше концентрируется на настройке времени обучения, когда IBM предоставляет лучший выбор алгоритмов обучения

(но для неопытного пользователя эта возможность менее важна, чем для математика). После нескольких часов расчета и обучения модели обе системы предоставляют итоговую модель (одну для Google Auto ML и до четырех для Watson Auto AI) с одинаковым набором показателей и диаграмм:

- Матрица ошибок
- Область ниже ROC кривой
- График ROC кривой
- Precision
- Accuracy
- Метрика F1

Числовые значения данных показателей сгруппированы в таблице ниже:

TABLE I. СРАВНЕНИЕ СТАТИСТИЧЕСКИХ ПАРАМЕТРОВ ОБУЧЕННЫХ МОДЕЛЕЙ

Метрики качества	**Платформа**	
	Google Auto ML	***IBM Auto AI***
Область под ROC кривой	0.774	0.914
Precision	77.8%	85.7%
Accuracy	83.3%	87.5%
F1 Measure	0.609	0.743

Несложно заметить, что Auto AI демонстрирует несколько лучшие результаты чем Auto ML.

Тестирование моделей. Т.к. обе платформы поддерживают опцию Batch Prediction, обработку загруженных вручную наборов данных через обученный AI-классификатор, для тестирования обученной модели были отобраны еще 500 вручную размеченных образцов и загружены единым .CSV файлом. Получив таблицу с теми же переменными и структурой, что и у обучающих выборок, обе платформы возвращают ту же таблицу данных, обогащенную данными о вероятной принадлежности каждого объекта к обоим классам. Тесты показали количество ложноположительных и ложноотрицательных результатов близки к тем, которые можно предсказать с помощью параметров Precision и Accuracy каждой модели, но с любопытным различием: Auto ML возвращает оценки вероятности с большим количеством значимых чисел (например, вероятность для Auto ML 82,3% из 100, а для AI - 0,8 из 1).

Заключение. Суммируя вышесказанное, можно сказать, что обе системы достаточно хороши для небольших социальных исследований. Google Auto ML более удобен для исследователей без сильного математического образования, распознает входные данные более качественно и дает более удовлетворительные результаты (по крайней мере, психологически), но работает медленнее и стоит больше.

С другой стороны, Watson Auto AI имеет тенденцию быть более доступным, быстрым и точным, но требует больше ручной работы и математических знаний. Обе платформы подходят для обработки результатов социально-экономических исследований без технических специалистов.

Благодарности. Исследование было выполнено при поддержке РФФИ (проект №18-37-00344)

ЛИТЕРАТУРА

1. Кенин А. М., Мазуров В. Д. Опыт применения нейронных сетей в экономических задачах [Электронный ресурс] URL: http://www.uralstars.com/Docs/Editor/Neuro.htm Доступ свободный

2. Васенков Д.В. Методы обучения искусственных нейронных сетей [Электронный ресурс] URL: http://www.aiportal.ru/articles/neural-networks/learning-neunet.html Доступ свободный

3. MehryarMohri, AfshinRostamizadeh, AmeetTalwalkar (2012) Foundations of Machine Learning, The MIT Press. 417 c.

4. G. James (2003) Variance and Bias for General Loss Functions, Machine Learning [Электронный ресурс] Free access URL: http://www-bcf.usc.edu/~gareth/research/bv.pdf Доступ свободный

5. Machine-learning-as-a-service Platforms comparison [Электронный ресурс] URL: https://www.altexsoft.com/blog/datascience/comparing-machine-learning-as-a-service-amazon-microsoft-azure-google-cloud-ai-ibm-watson/ Доступ свободный

6. Google AutoML official documentation page [Электронный ресурс] URL: https://cloud.google.com/automl-tables/docs/beginners-guide Доступ свободный

7. IBM Auto AI official documentation page [Электронный ресурс] URL: https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/ml-overview.html?linkInPage=true

8. CSV open format specification [Электронный ресурс] URL: https://arquivo.pt/wayback/20160521044400/http://mastpoint.curzonnassau.com/csv-1203/ Доступ свободный

СРАВНИТЕЛЬНАЯ ХАРАКТЕРИСТИКА СРЕД РАЗРАБОТКИ МОБИЛЬНЫХ ПРИЛОЖЕНИЙ ПОД ANDROID

С.М. Савченко, А.Э. Евстафиевская
(г. Томск, Томский политехнический университет)
sms14@tpu.ru, aee5@tpu.ru

COMPARISON OF ANDROID MOBILE APPLICATION DEVELOPMENT ENVIRONMENTS

S.M.Savchenko, A.E.Evstafievskaya
(Tomsk, Tomsk Polytechnik University)

Abstract. This article provides a comparative description of the most popular mobile application development environments. In total, article 5 sections, including introduction and conclusion.

Keywords. Eclipse, Android, Xamarin, IDE, Application.

Введение. В наше время трудно представить человека без мобильного телефона. В современном мире он необходим не только для коммуникации и связи с людьми, находящимися на расстоянии, но и для выполнения различных вычислительных, информационных и других функций.

В связи с такими высокими требованиями к мобильным устройствам возникает необходимость в разработке мобильных приложений, позволяющих реализовать все необходимые задачи.

В данной статье описаны три среды разработки мобильных приложений, а именно *Eclipse, Android Studio* и *VS Xamarin.*

Eclipse. *Eclipse* – это расширяемая *IDE* (интегрированная среда разработки). *IDE* – удобно организованный набор инструментов, необходимых для работы над программным проектом.

Eclipse – универсальная платформа, которая может использоваться для разработки приложений на любом языке программирования, но изначально «родным» для *Eclipse* является *Java* (на которой, кстати, сам *Eclipse* и написан).