УДК 004.24:004.6

АНАЛИЗ ЭФФЕКТИВНОСТИ МАШИННЫХ АЛГОРИТМОВ ПРИ РАБОТЕ С БОЛЬШИМИ ОБЪЕМАМИ ДАННЫХ

А.В. Титова
Магистрантка кафедры инженерной психологии и эргономики БГУИР

К.Д. Яшин
Заведующий кафедрой инженерной психологии и эргономики БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: kafipie@bsuir.by

Аннотация. Представлен начальный опыт участия магистрантки кафедры инженерной психологии и эргономики БГУИР в совместном проекте с лабораторией IBM (Канада) и профессором Чикагского университета (США) по анализу эффективности алгоритмов машинного обучения при работе с большими объемами данных.

Ключевые слова: рекомендатор, корреляционный анализ, бенчмаркинг, PySpark.

А.В. Титова
*В 2017 году окончила факультет прикладной математики и информатики БГУ, специальность «Компьютерная безопасность». Квалификация математик, специалист по компьютерной безопасности. Ассистент кафедры физико-математических дисциплин ИИТ БГУИР***.** *Магистрантка кафедры инженерной психологии и эргономики БГУИР.*

К.Д. Яшин
Руководитель научно-исследовательской группы НИГ – 7.1 «Системы и приборы экологического мониторинга в управлении безопасностью жизнедеятельности». Один из организаторов конференции международной научно-практической конференции BIG DATA and Advanced Analytics.

Введение

Для подготовки в Белорусском государственном университете информатики и радиоэлектроники специалистов по анализу больших данных кафедра инженерной психологии и эргономики, совместно с Чикагским университетом (США) и лабораторией IBM (Канада), приняла участие в научно-исследовательском проекте. Проект направлен на решение проблемы анализа эффективности различных алгоритмов машинного обучения для широкого спектра задач разработки программного обеспечения.

Целью проекта является анализ характеристик производительности реализованных в библиотеках Python алгоритмов корреляционного анализа.

Корреляционный анализ статистических данных широко применяется в экономике, астрофизике и социальных науках (в частности в психологии и социологии) [1]. Сфера применения корреляционного анализа обширна: контроль качества промышленной продукции,

металловедение, агрохимия, гидробиология, биометрия и проч. В прикладных направлениях приняты разные границы интервалов для оценки плотности и значимости связей технологических параметров. При этом выбор статистической модели представляет собой трудоемкий процесс из-за сложностей, как в алгоритмах, так и в структуре данных. Поскольку многие компании начинают более активно использовать аналитику для развития своих бизнес-процессов, значение выбора статистических моделей возрастает. Настоящее исследование направлено на выбор модели путем разработки автоматизированного программного комплекса (рекомендатора) для получения оптимального алгоритма машинного обучения из библиотеки машинного обучения PySpark. Работа над проектом и решение поставленных задач выполнялись по схеме (рисунок 1).

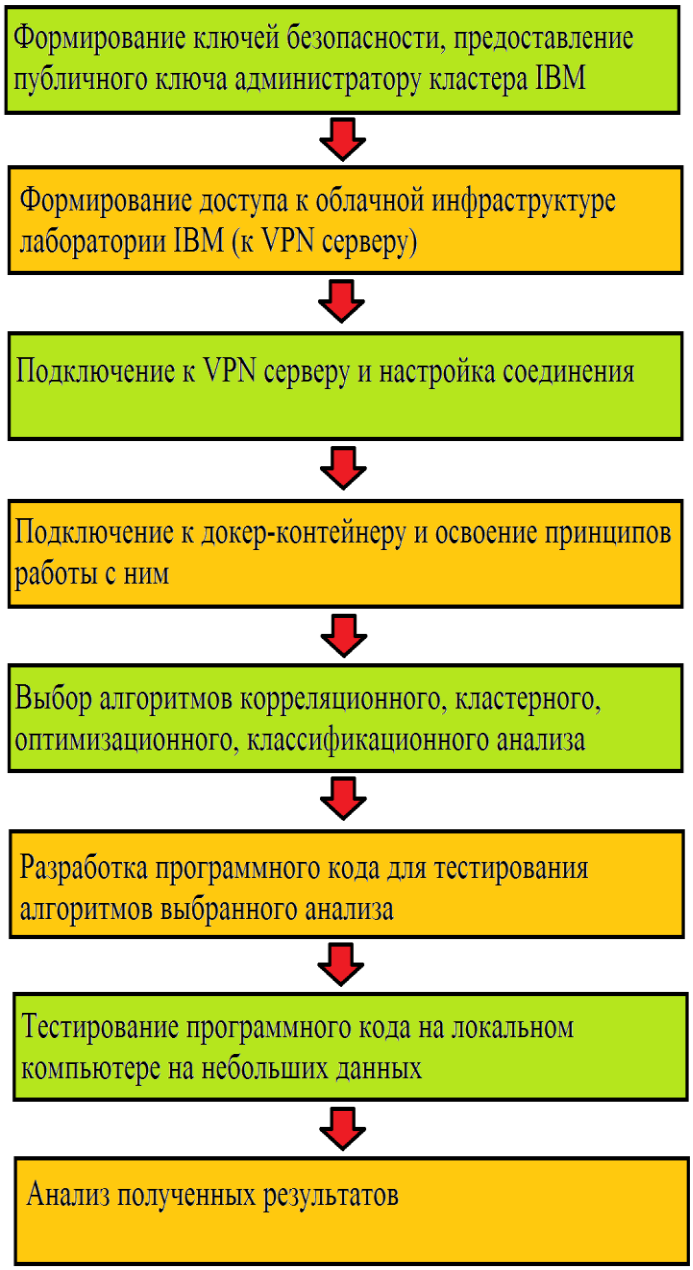
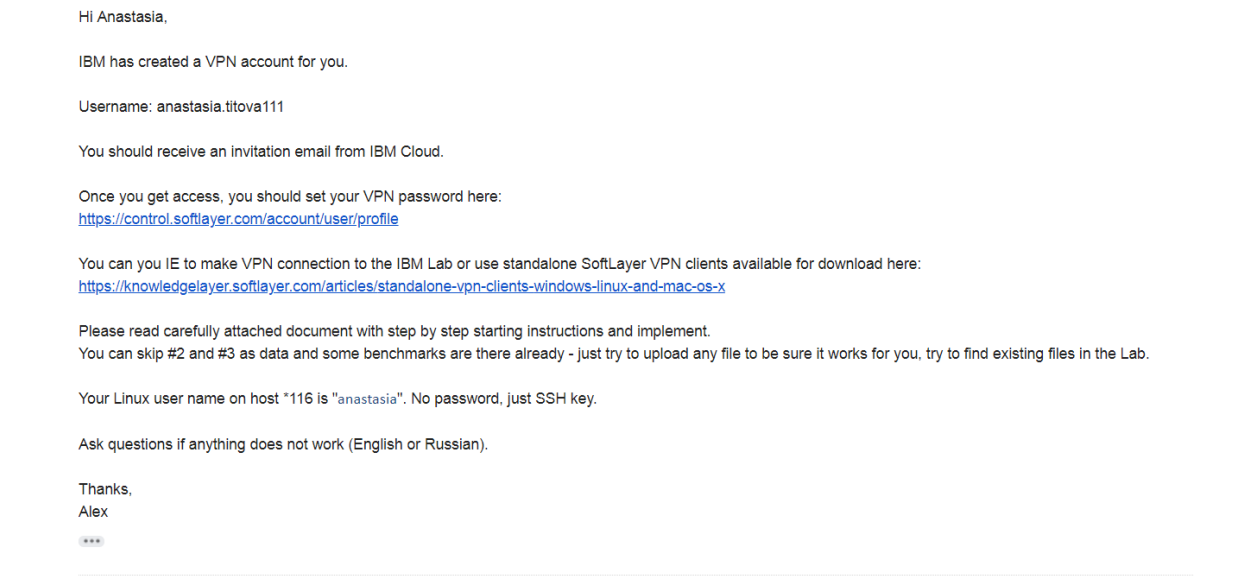


Рисунок 1. Схема алгоритма выполнения проекта

Результаты

Получение доступа к облачной лаборатории IBM. Регистрация и получение доступа к облачной лаборатории IBM, настройка доступа к VPN серверу с помощью сгенерированного и отправленного администратору ранее публичного ключа шифрования занимает от нескольких недель до нескольких месяцев, в зависимости от загруженности администраторов сервиса. На рисунках 2-4 продемонстрированы шаги по получению доступа к виртуальной лаборатории IBM.



Hi Anastasia,

IBM has created a VPN account for you.

Username: anastasia.titova111

You should receive an invitation email from IBM Cloud.

Once you get access, you should set your VPN password here:
https://control.softlayer.com/account/user/profile

You can you IE to make VPN connection to the IBM Lab or use standalone SoftLayer VPN clients available for download here:
https://knowledgelayer.softlayer.com/articles/standalone-vpn-clients-windows-linux-and-mac-os-x

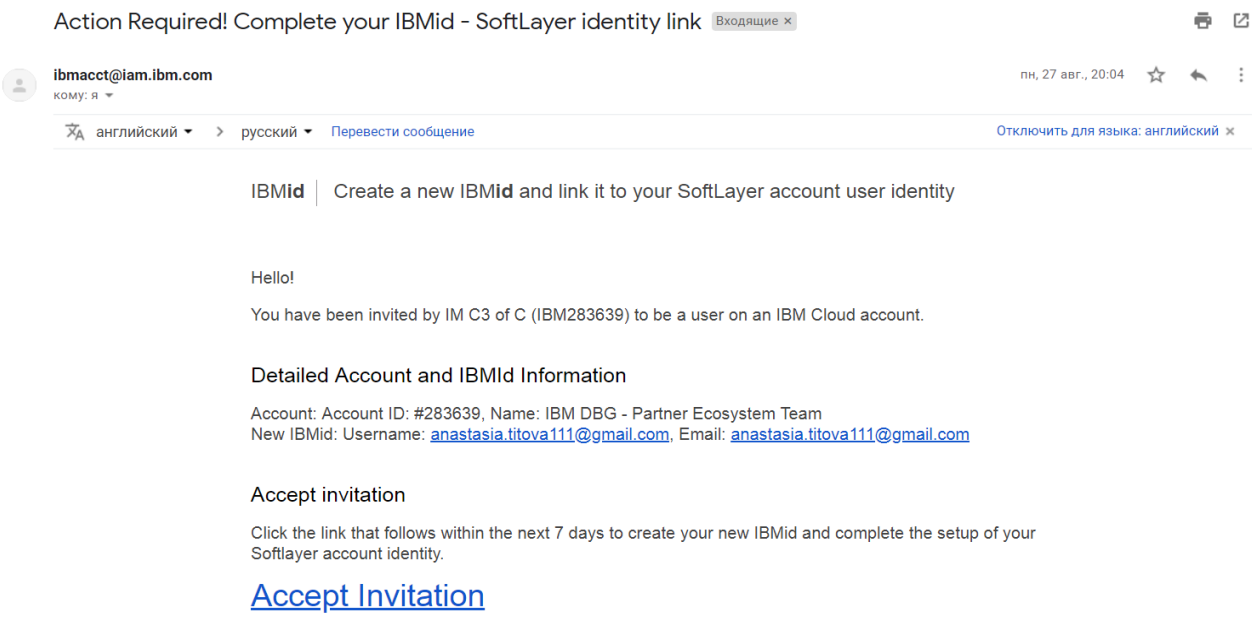
Please read carefully attached document with step by step starting instructions and implement.
You can skip #2 and #3 as data and some benchmarks are there already - just try to upload any file to be sure it works for you, try to find existing files in the Lab.

Your Linux user name on host "116 is "anastasia". No password, just SSH key.

Ask questions if anything does not work (English or Russian).

Thanks,
Alex

Рисунок 2. Сообщение о создании аккаунта от куратора проекта из Чикагского университета



Action Required! Complete your IBMid - SoftLayer identity link Входящие ×

ibmacct@iam.ibm.com
кому: я

пн, 27 авг., 20:04

английский › русский Перевести сообщение Отключить для языка: английский ×

IBMid | Create a new IBMid and link it to your SoftLayer account user identity

Hello!

You have been invited by IM C3 of C (IBM283639) to be a user on an IBM Cloud account.

Detailed Account and IBMId Information

Account: Account ID: #283639, Name: IBM DBG - Partner Ecosystem Team
New IBMid: Username: anastasia.titova111@gmail.com, Email: anastasia.titova111@gmail.com

Accept invitation

Click the link that follows within the next 7 days to create your new IBMid and complete the setup of your Softlayer account identity.

Accept Invitation

Рисунок 3. Сообщение о регистрации в лаборатории IBM

Пара ключей генерировалась с помощью программы PuttyGen. Публичный ключ пересылается администратору лаборатории IBM для создания личной учетной записи. Личный ключ использовался для подтверждения аутентификации на VPN сервере.

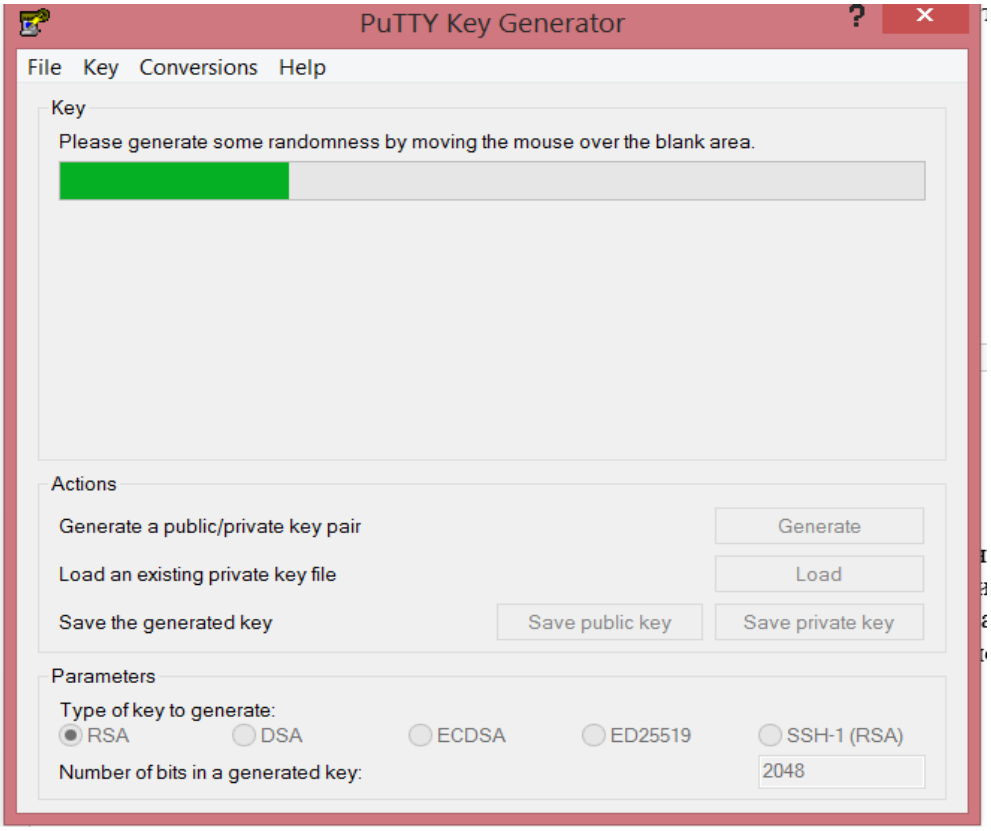


Рисунок 4. Интерфейс программы PuttyGen при генерация ключей

Работа с докер-контейнером. После создания учетной записи осуществляется подключение к VPN серверу. Далее выполняется загрузка кода программы в докер-контейнер. Докер – технология, позволяющая запускать приложение в контейнере, похожем на виртуальную машину. Основные черты, отличающие его от виртуальной машины, – это легковесность, ресурсоёмкость и практически полная независимость от инфраструктуры [2].

Подключение в VPN серверу. На рисунках 5-7 представлены шаги подключения к VPN серверу. 1) Открытие соединения с VPN-сервером производилось через сайт [3] (работает только в Internet Explorer). 2) Учетная запись привязана к соответствующему личному кабинету IBM кластера. 3) При подключении пользователю автоматически предлагается скачать программу для соединения, после установки которой и удачного подключения появляется сообщение.

Настройка соединения с контейнером. Для соединения с контейнером требуется скачать программу Putty и выполнить действия, проиллюстрированные на рисунках 8-11. 1) Вводится IP-адрес. 2) Во вкладке SSH -> Auth на боковой панели загружается личный ключ из пары ключей, которые были сгенерированы для получения доступа. 3) Следует возврат на вкладку Session и сохранение подключения под своим именем. Программа предложит ввести логин. 4) После аутентификации появится сообщение об открытии соединения с контейнером.

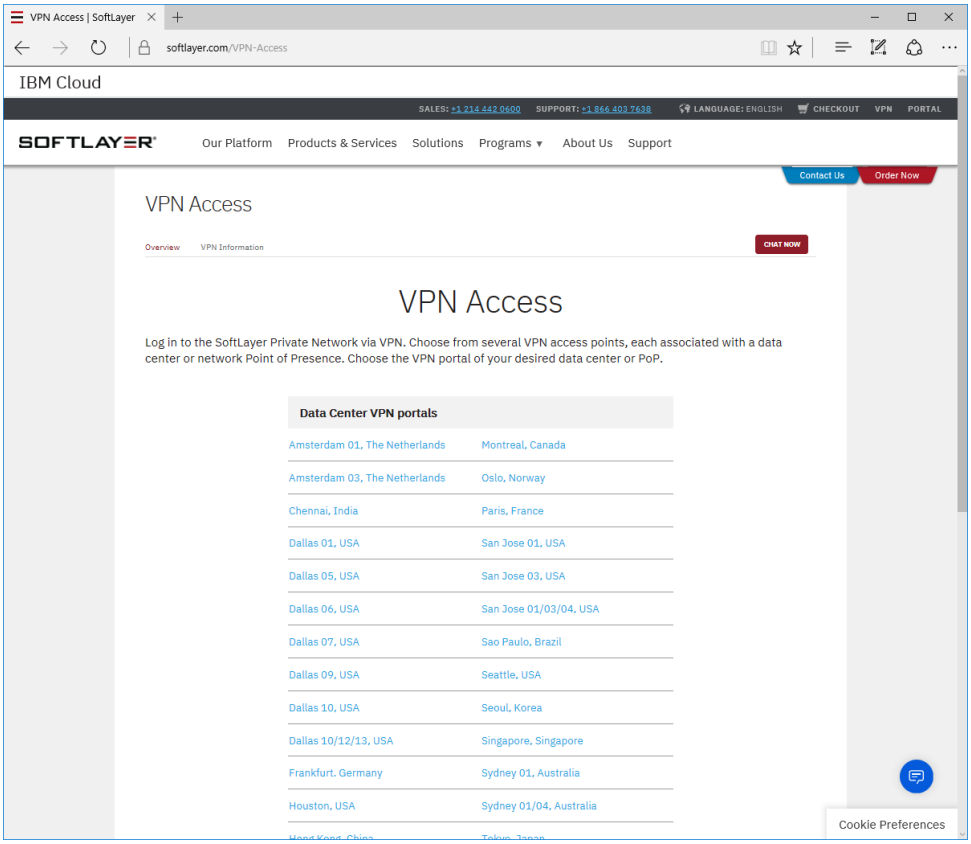


Рисунок 5. Список доступных VPN серверов

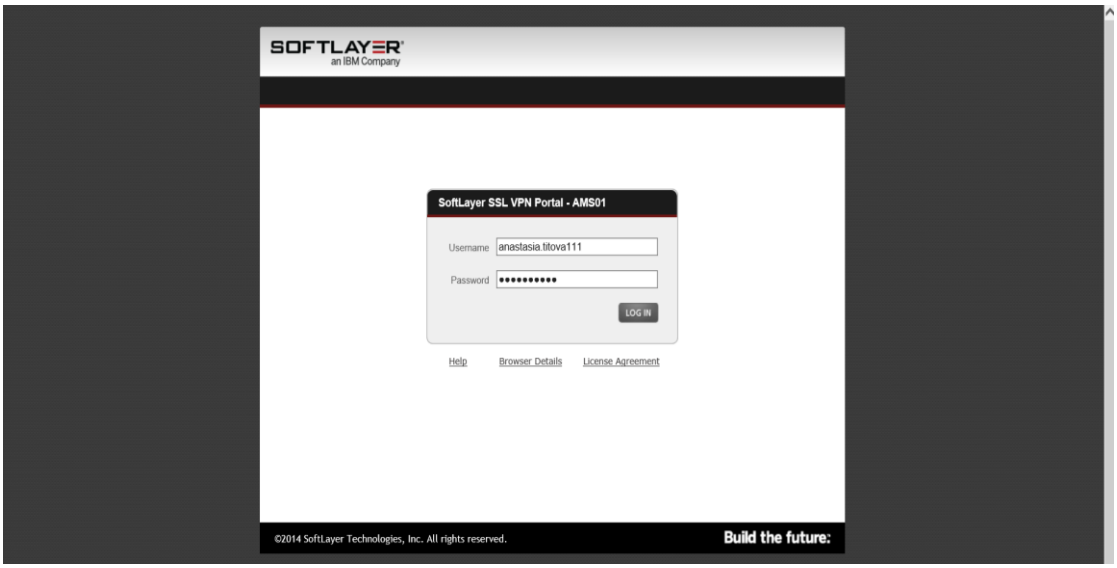


Рисунок 6. Окно подключения к серверу с помощью аутентификационных данных IBM

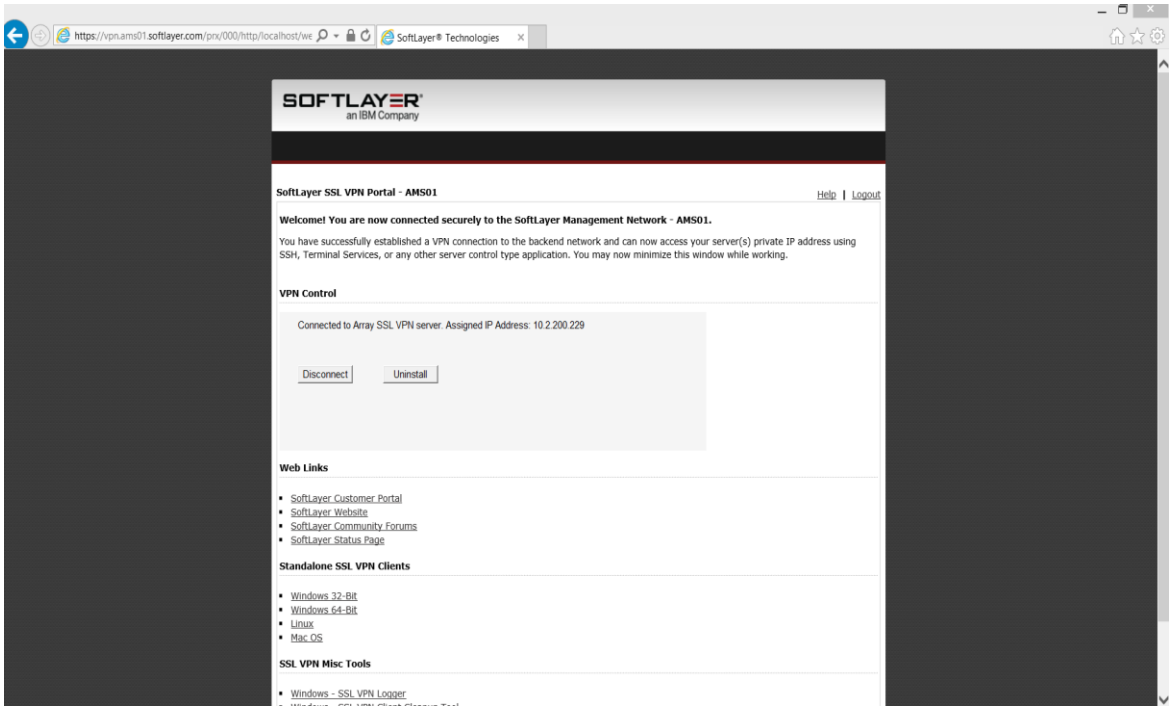


Рисунок 7. Окно установления соединения с сервером

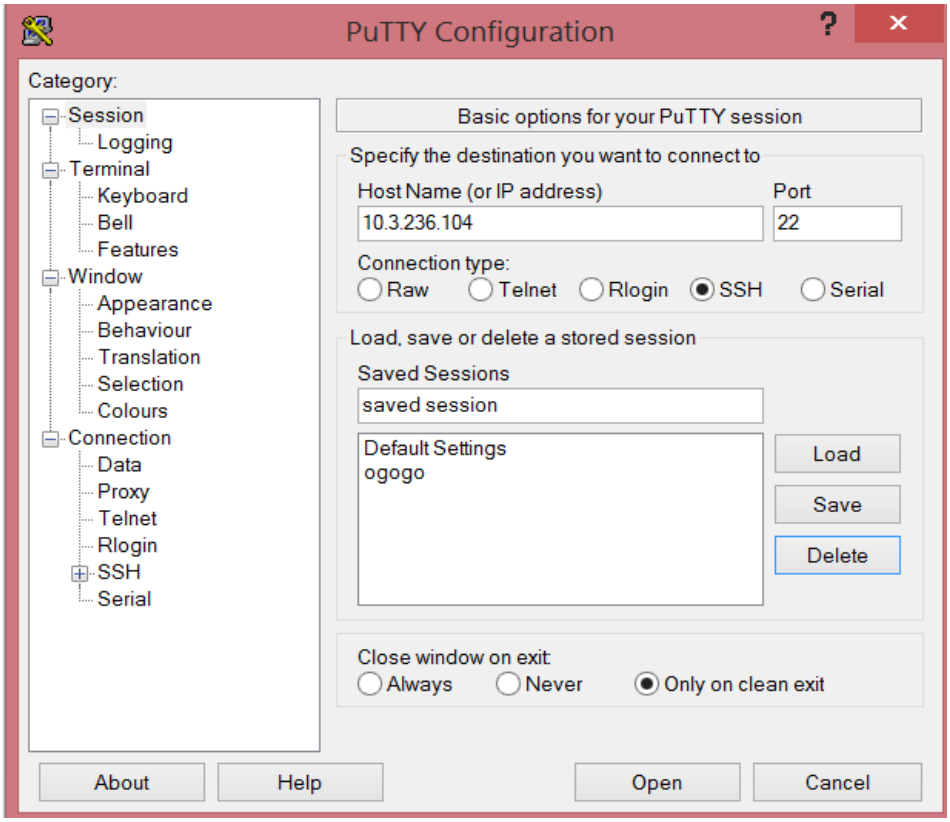


Рисунок 8. Окно настройки сессии

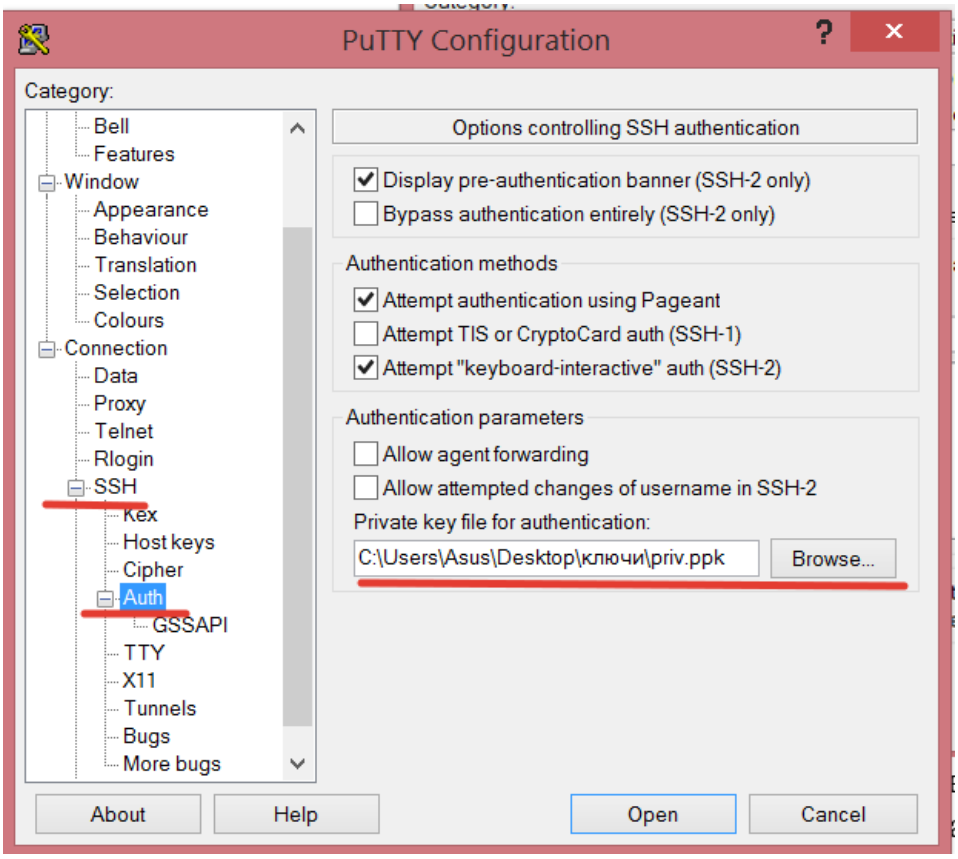


Рисунок 9. Окно подключения личного ключа безопасности

Рисунок 10. Окно соединения с контейнером

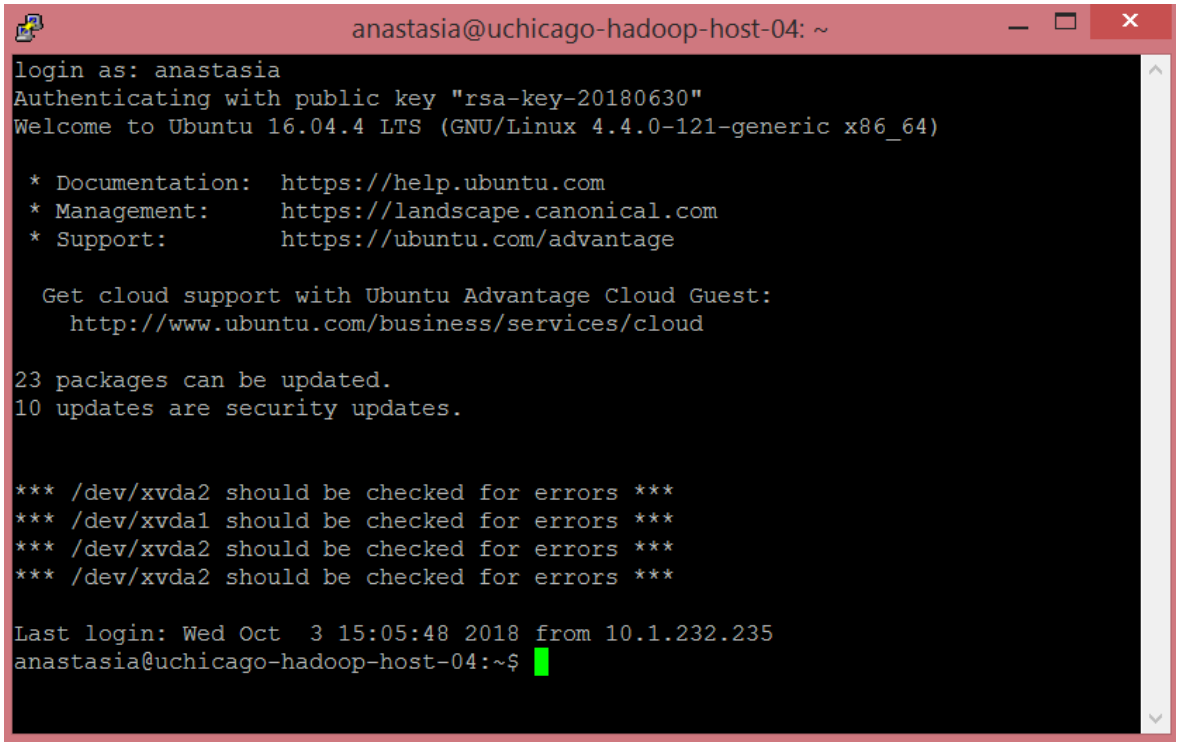


Рисунок 11. Окно открытого докер-контейнера

В ходе выполнения проекта возникли некоторые трудности при работе с кластером IBM. В ожидании решения вопроса с докер-контейнером проекта участникам эксперимента было предложено уделить внимание разработке компьютерных программ для реализации алгоритмов корреляционного, оптимизационного, кластерного, классификационного анализов. Кроме того, участникам эксперимента было предложено уделить внимание тестированию самостоятельно разработанных участниками эксперимента алгоритмов. Эксперимент продолжался с применением и обработкой информационных данных, локализованных на персональном компьютере.

Для группы исследователей кафедры инженерной психологии и эргономики в качестве предмета исследования выбран корреляционный анализ. Изучена теория алгоритмов корреляционного анализа в библиотеках Python. Выбраны следующие алгоритмы.

Коэффициент корреляции Пирсона [4]

$$\rho_{X,\,Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Пример программной реализации для Python в библиотеке pandas [4]:

```
odel_year', 'origin'], axis=1).corr(method='pearson')
```

Коэффициент корреляции Спирмана [4]

$$\rho_{X,\,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Пример программной реализации для Python в библиотеке pandas [4]:

```
mpg_data.drop(['model_year', 'origin'], axis=1).corr(method='spearman')
```

Группа кафедры инженерной психологии и эргономики выбрала для исследования два указанных выше вида корреляционных зависимостей, предполагая в дальнейшем взять еще несколько алгоритмов для изучения. Далее основной задачей явилось установление различий в эффективности работы алгоритмов. Это осуществлялось путем тестирования алгоритмов на наборах информационных данных. Последние отличаются как по размерам, так и по структуре. На этапе эксперимента предполагается, что все информационные данные, используемые для тестов эффективности, уже обработаны и не имеют структурных ошибок, таких как отсутствие значений, значения неверного формата и т.д.

Всем группам, участвующим в проекте, был открыт доступ к материалам проектов лаборатории IBM на Dropbox (рисунок 12). Среди этих материалов имеется исходный Python-код для алгоритмов регрессионного анализа. Кроме того, имеется 4 пары наборов информационных данных, представленных в формате .csv (рисунок 13). Эти информационные данные можно использовать на следующих шагах проекта для тестирования алгоритмов, разработанных авторами проекта. Также были предоставлены реализованные в RStudio модели оценки эффективности алгоритмов по некоторым 5 параметрам (таблица 1), а также инструкции с описанием необходимых настроек и программа рекомендатора на языке R.

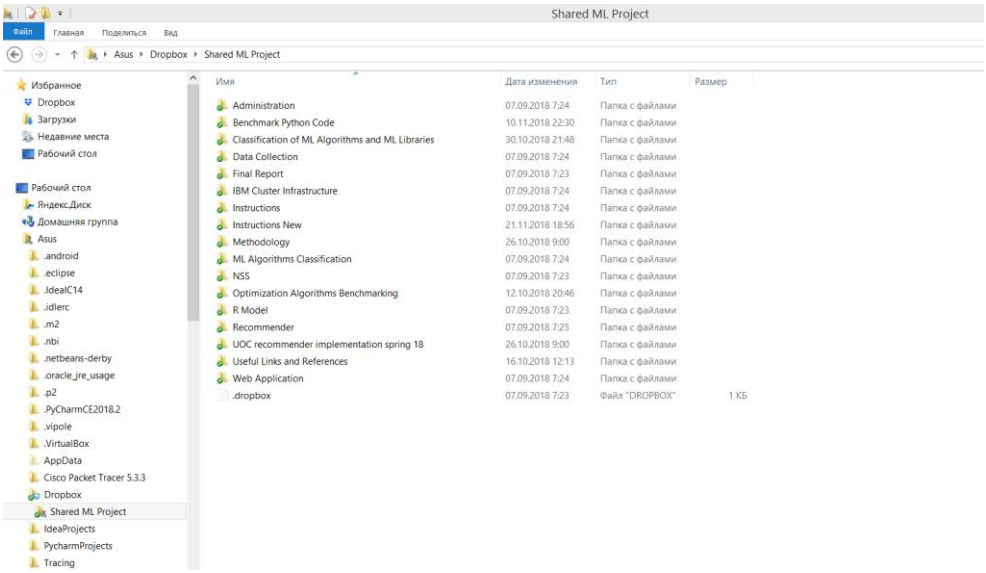


Рисунок 12. Результаты учебных проектов лаборатории IBM

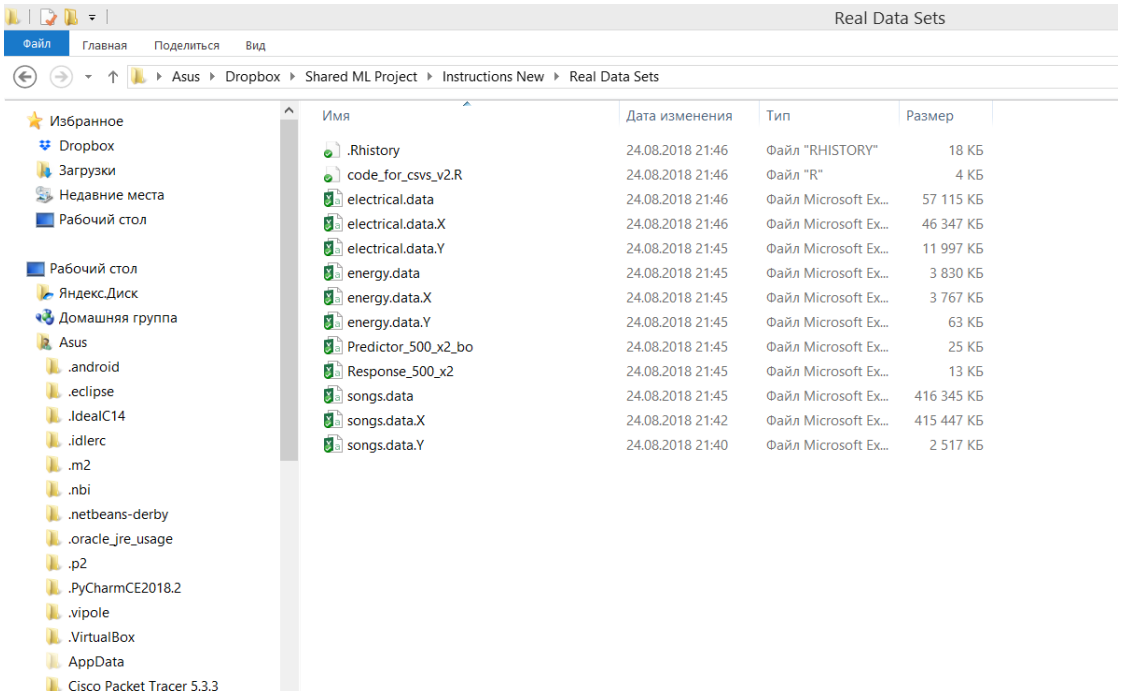


Рисунок 13. Наборы данных для тестирования

Процесс бенчмаркинга собирает и анализирует информационные данные по характеристикам производительности. Рекомендатор содержит пять вложенных статистических моделей измерения производительности. Они прогнозируют объем компьютерных ресурсов, требуемых каждым алгоритмом. Это позволит пользователям рекомендатора изменять размеры и вес набора информационных данных по пяти показателям производительности, учитывая бизнес-требования. Программа рекомендатор генерирует оценку для каждого из алгоритмов.

Таблица 1

Список характеристик производительности [5]

Характеристика	Определение
Точность	RMSE используется для измерения доли дисперсии в зависимой переменной. Минимальная RMSE соответствует оптимальной модели
Время отклика	Время, требующееся алгоритму для вывода результата
Скорость чтений физического блока	Количество блоков данных, считанных с диска в секунду, при заданном алгоритме
Использование CPU	Использование компьютерных ресурсов обработки при запуске указанного алгоритма
Количество операций ввода/вывода	Общее количество физических чтений и записей, которые имеют место при заданном алгоритме

Список файлов статистических моделей оценки эффективности, используемых рекомендатором, представлен на рисунке 14. Их использование оправдано на этапе работы с большими данными.

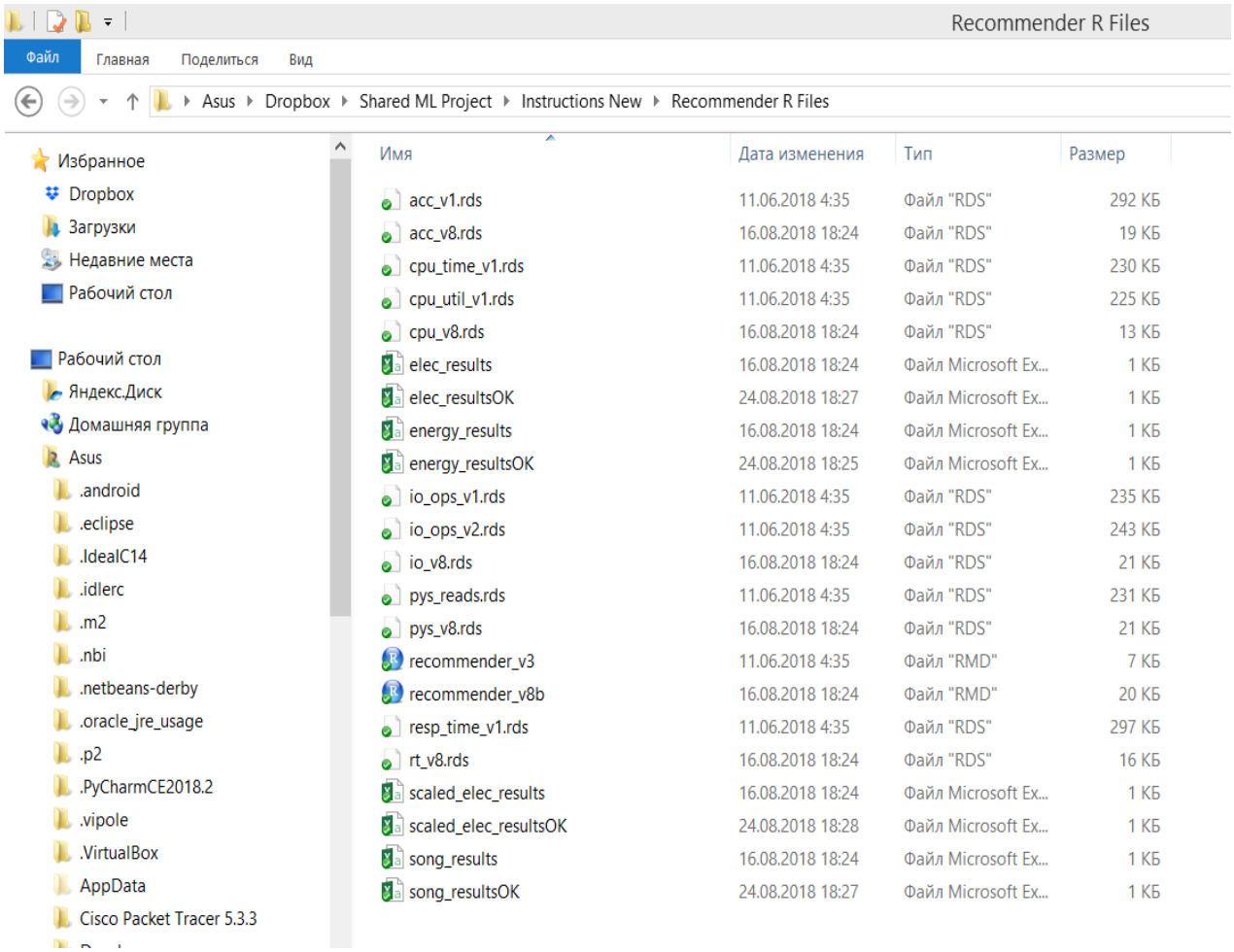


Рисунок 14. Список файлов программной реализации рекомендатора

В каждую из программ, реализующих модели оценки эффективности алгоритмов, направляются результаты исполнения алгоритма на Python в формате .csv. Результаты оцениваются по соответствующему параметру эффективности. Затем осуществляется процесс визуализации этих данных в виде графиков в RStudio. На рисунке 15 представлены зависимости точности алгоритмов от количества предикторов только для первой программы accuracy_v8; для остальных четырех программ получены похожие графики. Предиктор – это параметр, по которому сравниваются данные. Видно, что точность резко возрастает для больших объемов данных при количестве предикторов, превышающем 2000.

На рисунке 16 представлена зависимость точности модели для различных алгоритмов от размера данных (ось obs). Алгоритм GLM наименее предпочтителен, если точность вычислений в приоретете.

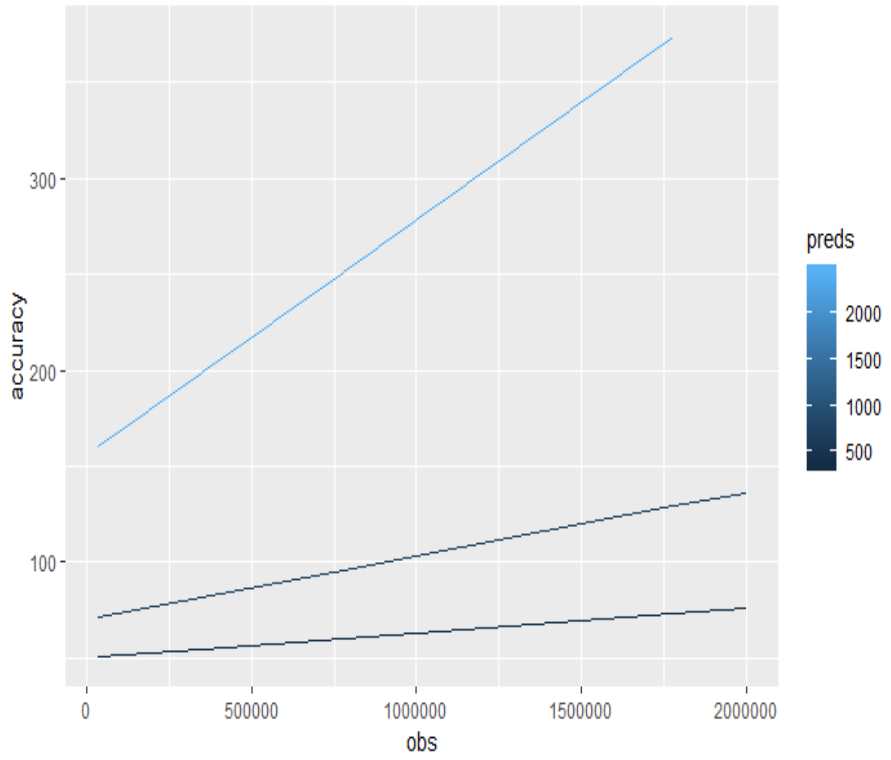


Рисунок 15. Зависимость точности алгоритмовв от количества предикторов

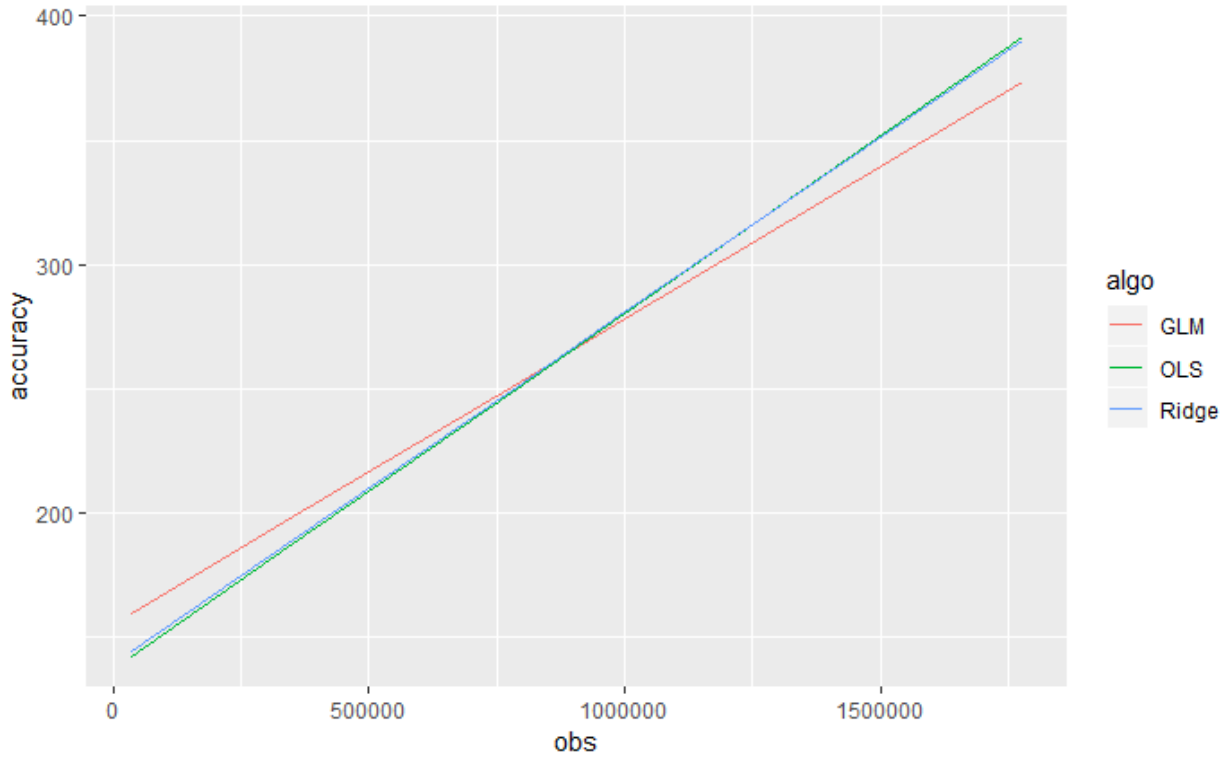


Рисунок 16. Зависимость точности алгоритмов от размеров данных

Заключение и перспективы развития

В работе рассмотрены вопросы актуальности анализа эффективности алгоритмов машинного обучения, используемых для задач бизнеса и промышленности. Представлено описание научно-исследовательского проекта, выполненного магистранткой кафедры инженерной психологии и эргономики под руководством профессора Чикагского университета (США) и специалистов лаборатории IBM (Канада). Представлена методология выполнения поставленных в проекте задач, описаны результаты выполнения проекта. Подробно описаны этапы настройки необходимой среды и начальные тестовые данные выбранных алгоритмов корреляционного анализа.

Развитие проекта направлено на освоение технологии обработки больших объемов информационных данных бизнеса и промышленности, а также на их глубокий анализ.

Авторы благодарят Бориса Зибицкера и Алекса Луперсольского за оказание технической помощи и консультаций при выполнении работы.

Литература

[1] Корреляция: Материал из Википедии — свободной энциклопедии: Версия 93206762, сохранённая в 10:05 UTC 10 июня 2018 / Википедия, свободная энциклопедия. — Электрон. дан. — Сан-Франциско: Фонд Викимедиа, 2018. — Режим доступа: https://ru.wikipedia.org/?oldid=93206762

[2] 2. Черных А. Основы работы с Docker [Электронный ресурс] / ООО «Селектел», Санкт-Петербург, 2016. – Режим доступа: https://community.vscale.io/hc/ru/community/posts/211783625-Основы-работы-с-Docker

[3] Руководство для подключения к VPN серверу [Электронный ресурс] / SoftLayer Technologies. Inc., United States, 2019. – Режим доступа: https://www.softlayer.com/VPN-Access

[4] 4. Далинина Р. Введение в корреляционный анализ [Электронный ресурс] / Oracle, United States, 2017. – Режим доступа: https://www.datascience.com/blog/introduction-to-correlation-learn-data-science-tutorials

[5] Okallau B., Shebik B., Wishart H. Recommender Development for Selection of Appropriate Regression Algorithms in Pyspark.ML Library (материал предоставлен профессором Чикагского университета Б. Зибицкером).

MACHINE ALGORITHMS EFFICIENCY ANALYSIS WORKING WITH BIG DATA

A. TITOVA
Master student of the department of engineering psychology and ergonomics BSUIR

K. YASHIN[2],PHD
Head of the Department of Human Engineering and Ergonomics, BSUIR

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: kafipie@bsuir.by

Abstract. The initial experience of the postgraduate student of the BSUIR Department of Engineering Psychology and Ergonomics participation in a joint project with the IBM laboratory (Canada) and a professor of the University in Chicago (USA) on analyzing the effectiveness of machine learning algorithms when working with big data is presented.

Keywords: recommender, correlation analysis, benchmarking, PySpark.