Using cloud-based machine learning technologies in limited funded social research

Romanchukov SV 1,5, Berestneva OG 2, Berezhnova EV 3,4
1 National Research Tomsk Polytechnic University, School of Computer Science & Robotics, Tomsk, Lenin ave. 30, 634050, Russia
2 National Research Tomsk Polytechnic University, School of Computer Science & Robotics, Tomsk, Lenin ave. 30, 634050, Russia
3 Moscow State Institute of International Relations (University) of the Russian Foreign Ministry, Moskow, Prospect Vernadskogo 76, 119454, Russia
4 Moscow state University of civil engineering (National research university), Moskow, Yaroslavskoye Shosse 26, 129337, Russia
5 inoytomsk@yandex.ru

Abstract. This work is devoted to the description of cloud solutions for machine learning, which are already used in the business data analysis and may be applicable in the social sciences. First of all, the article is addressed to specialists in sociology/psychology/economics/gender studies, who need a deep analysis of the accumulated data, but at the same time do not have sufficient expertise in the field of mathematics, machine learning and Big Data processing and/or do not have sufficient funding to support the staff of professional analysts or data scientists. Using as an example one dataset, the size and structure of which are comparable with those for various social studies, we go through all stages of training and testing the model in Google Cloud AI and IBM Watson Auto AI, comparing their advantages and disadvantages.

1. Introduction

Machine learning and Artificial neural networks are quite effective in solving problems of classification, pattern recognition, complex systems behavior prediction and selecting unknown parameters that relate the characteristics of complex objects, including economic systems [1]. There are several methods of so-called Neural network “training”: learning with a teacher, learning without a teacher, learning with reinforcement [2].

The first option - learning with a teacher, a machine learning method in which the test system is forcibly trained using examples of "stimulus-response" - seems to be the most suitable for solving the problem of the relationship between the two groups of parameters, classification or parameter prediction by given evidences [3].

The method of education with a teacher suggests that there may be some relationship between the inputs and the reference outputs, but it is unknown. Only the final set of precedents is known - the stimulus-response pair, called the training set. On the basis of these data, an iterative process of selecting parameters for the purpose of restoring dependence and constructing a relationship model suitable for forecasting, capable of giving a fairly accurate answer for any object, occurs. To measure the accuracy of answers, as well as in learning by examples, a quality functional can be introduced [4].

Organizing the learning process from scratch is a non-trivial mathematical and programming task, however, at the moment, a significant number of cloud software products are available on the market that allow you to create AI and ML projects without special knowledge. Research groups with limited
funding and no access to professional expertise in data science and programming can use such software solutions. To be suitable for such audience software product should meet several additional requirements other than precision and technical parameters:

- Clear and easy to learn documentation / interface
- Ease of use and results interpretation
- Cost and features available for a minimal price or in trial set

As the initial data array on which the systems were compared, a sample of companies was taken, including data from January 1, 2018 to September 2019 - 1640 data rows and 24 variables describing their commercial indicators, the parameters of their advertising campaigns and internal social indicators related to labor disciplines, including but not limited to:

- Company name and URL
- Industry and company type
- ARR
- Main advertising channel
- Ad spends (evaluated by SEMRush and SPYfu separately)
- Moat creatives count
- Ghostery terms
- Employee count, employee increase and summary of anonymous responses about working conditions/atmosphere

This set of numerical and categorical variables, showing both economical and social policy within the company, was used to predict one logical variable (with only "True" or "False" variants) - do this company perform one specific action or not. That is basically a task of binary classification - that can be solved by means of Machine Learning algorithms and can serve a good testing field to compare different ones.

2. Machine learning as a service

First, we need to introduce Machine learning as a service (MLaaS) approach already used for marketing research and operating Big Data in different business structures, but yet underestimated in less commercial-oriented social studies.

MLaaS is an umbrella definition of various cloud-based platforms that cover data pre-processing, model training, and model evaluation, with further prediction. Prediction results can be bridged with user's internal IT infrastructure through REST API or returned via a user interface of the platform itself. There is a number of MLaaS platforms that allow for fast model training and deployment available, but the four leading cloud MLaaS services are [5]:

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

In this paper we will concentrate on the first two platforms, basing on assumption that Google and IBM products are closer to software already widespread in social studies (e.g. Google Forms, Google Sheets, IBM SPSS) - and can be even integrated with data processed via this software.

Both platforms utilize very similar Machine Learning workflow ("standard machine learning workflow") [6][7]:

- Data preparation: Make sure your data is properly formatted before and after data import
- Model Train: Set parameters and build your model
- Model evaluation by a set of model metrics provided: at this stage user can compare several models trained differently
- Model testing: prediction on a set of test data with pre-known outcomes
- Model deployment and prediction that makes a trained and tested model finally ready-for-use.

Both systems provide similar options (in term of machine learning components):

- Binary classification model,
- Multi-class classification model,
-...

2
Regression model for a continuous value. Watson provides slightly more options for choosing solving algorithms and their parameters, but Google have more user-friendly user interface (UI) than Watson ML - with highly detailed support and information features. It requires less expertise to operate, mostly because of better implemented right in the UI. Almost any element, term or chart have a button with a "question" sign near it giving short explanation about this element's purpose and math behind it.

Watson Auto ML provides similar components without any explanation. This is not a serious issue for a data scientist or mathematician, but researcher less experienced in Data Science and more concentrated on sociology or gender studies, for example, would find first variant more supporting, e.g. comparing two models by "Area Under ROC Curve" is senseless without the knowledge what "ROC" stands for. Google Auto ML tends to be easier to use for researchers without a strong mathematical background.

3. Platform Pricing
Both platforms base their pricing on amount of time and number of nodes needed to perform the task, and provide

3.1. Google Auto ML Tables
Google AutoML Tables provides a free trial by giving 6 free node hours each for training and for batch prediction, per billing account. After that, Google provide pricing based on type of operation:
  - Model training costs $19.32 per hour of compute resources
  - Model deployment costs $0.005 per GB per hour per machine (Auto ML replicate models to 9 machines for low latency serving purposes, so final price is 9 time bigger)
  - Batch prediction costs $1.16 per hour of compute resources used
  - Online predictions cost $0.21 per hour of compute resources used.

3.2. IBM Watson Auto AI
Watson pricing model is based on so-called capacity unit hours (or CUH) approach that allows to match different types of hardware. For example, 4 hours of using one NVIDIA K80, one hour of NVIDIA V100 GPU, or one hour of Auto AI prediction with 16 vCPU and 64 GB RAM machine costs equal amount of capacity unit hours.

The lite plan instance of the IBM Watson Machine Learning service is totally free and provides a maximum of five deployed models, 5,000 predictions per month, and 50 capacity unit hours per month. In case if this is not enough, Lite Plan may be upgraded to Standard that charges flat rate per each thousand of predictions and per total number of capacity unit-hours, with a minimum of 1 minute per training job - $0.54 USD per capacity unit-hour or 1000 predictions.

Overall testing of both products on the same dataset resulted in 1.81 h of model training on Google Platform (which is 30% of total trial amount of resources or ~35 USD of paid one), on Watson platform the same tests took 26% of CUH available for free each month. This allow us to conclude that on relatively small datasets typical for small research team Watson Auto AI tends to be more affordable.

4. Data preparation
Both system supports different ways of data ingestion:
  - From their own data infrastructure (Google Cloud/IBM Cloud)
  - From various integrated data sources and databases
  - From personal computer in form of a separate file

To imitate the most possible variant for a small research team or a student's project we ingested data from a PC using a .CVS format - one of the most popular file format for most of the table-based analytic software[8].

Both platforms support CSV-import and offers an automatic variable type recognition by default. Both platforms uploaded the data fast enough, but:
  - On the one hand Google product processed data almost 5 times slower (2.5 minutes against less than 30 seconds for Watson),
• On the other hand IBM Watson failed to recognize two categorical variables with high amount of categories and marked them as unique text strings. Integrated component for manual data-preparation and improvement is available for both this platforms, so this isn’t a serious issue, but it requires a researcher to finalize the process program was supposed to perform and eliminates the difference in processing speed. Five tests on random data samples were committed and resulted in the same situation: 

*Google loads data slower than Watson, but recognizes it more consistently*

5. Model Train and evaluation

After loading data and selecting a goal for education both solutions provide a set of additional options. Google is more concentrated on tweaking education time, when IBM provides a better selection of education algorithms (but for non-experienced user this possibility is less important that for a mathematician).

After hours of calculation and model education are finished, both systems provide resulting model (one for Google Auto ML and up to four for Watson Auto AI) with the same set of metrics and charts:

- A confusion matrix
- Area Under ROC Curve
- ROC Curve graph
- Precision
- Accuracy
- F1 Measure

All numerical variables for our testing data set are presented at the table below. Several tests with randomly generated samples resulted in the similar qualitative picture.

| Learning Estimators | ML platform | Google Auto ML | IBM Auto AI |
|---------------------|-------------|----------------|-------------|
| Area Under ROC      | 0.774       | 0.914          |
| Precision           | 77.8%       | 85.7%          |
| Accuracy            | 83.3%       | 87.5%          |
| F1 Measure          | 0.609       | 0.743          |

All metrics on the same set of data are higher for Auto AI than for Auto ML. 

*At this scope IBM algorythms look more precise, but this can be just a formal result of “over-education” or some kind of bias.*

6. Model testing

For model testing were selected 500 more samples with known outcome. Both platforms support Batch Prediction option - processing of manually loaded data sets via trained AI classifier. Receiving a table with the same variables and structure as the training samples had, they return the same data table enriched with the same notes and data of both class probabilities.

Tests showed amount of false positives and false negatives, that are close to those that can be predicted by given model's Precision and Accuracy, but with an interesting difference: 

*Auto ML provided possibilities with bigger amount of significant numbers (e.g. probability 82.3% for Auto ML is just 0.8 for Auto AI).*

7. Conclusion

Summarizing sated above, we can tell that both systems are quite good for small social studies. Google Auto ML tends to be easier to use for researchers without a strong mathematical background, recognizes input data more consistently and gives more satisfying results (at least psychologically), but is slower and more expensive.
On the other hand Watson Auto AI tends to be more affordable, faster and more precise, but requires more manual work and mathematical expertise. Both of platforms are suitable for processing results of social and economic studies without technicians.

Acknowledgments
The work was supported by the RFBR, and conducted in frames of project 18-37-00344 "Regional Development Social and Economic Parameter's Interdependence Analysis and Modeling" and partially supported by project 18-07-00543 “Intelligent support of managerial decision-making on innovative development of regional scientific medical centers”

References
[1] Kenin A M, Mazurov V D 2018 The experience of using neural networks in economic tasks] (in Russian: Opyt primeneniya neyronnykh setey v ekonomicheskikh zadachakh). URL: http://www.uralstars.com/Docs/Editor/Neuro.htm
[2] Vasenkov D V 2018 Methods of teaching artificial neural networks. URL: http://www.aiportal.ru/articles/neural-networks/learning-neunet.html
[3] Mohri M, Rostamizadeh A, Talwalkar A 2012 Foundations of Machine Learning, (Boston, MS: MIT Press).
[4] James G 2003 Variance and Bias for General Loss Functions, Machine Learning. URL: http://www-bcf.usc.edu/~gareth/research/bv.pdf
[5] Machine-learning-as-a-service 2016 Platforms comparison. URL: https://www.alterxsoft.com/blog/data science/comparing-machine-learning-as-a-service-amazon-microsoft-azure-google-cloud-ai-ibm-watson/
[6] Google AutoML official documentation page. 2019 URL: https://cloud.google.com/automl-tables/docs/beginners-guide
[7] IBM Auto AI official documentation page. 2019 URL: https://dataplatf orm.cloud.ibm.com/docs/content/wsj/analyze-data/ml-overview.html?linkInPage=true
[8] CSV open format specification. 2019 URL: https://arquivo.pt/wayback/20160521044400/http://mastpoint.curzonassau.com/csv-1203/