Machine Learning Algorithms for Food Intelligence: Towards a Method for More Accurate Predictions

Ioanna Polychronou(✉), Panagis Katsivelis, Mihalis Papakonstantinou, Giannis Stoitsis, and Nikos Manouselis

Agroknow, 110 Pentelis, 15126 Maroussi, Greece
{ioanna.polyxronou,katsivelis.panagis,mihalis.papakonstadinou, stoitsis,nikosm}@agroknow.com
https://www.agroknow.com

Abstract. It is evident that machine learning algorithms are being widely impacting industrial applications and platforms. Beyond typical research experimentation scenarios, there is a need for companies that wish to enhance their online data and analytics solutions to incorporate ways in which they can select, experiment, benchmark, parameterise and choose the version of a machine learning algorithm that seems to be most appropriate for their specific application context. In this paper, we describe such a need for a big data platform that supports food data analytics and intelligence. More specifically, we introduce Agroknow's big data platform and identify the need to extend it with a flexible and interactive experimentation environment where different machine learning algorithms can be tested using a variation of synthetic and real data. A typical usage scenario is described, based on our need to experiment with various machine learning algorithms to support price prediction for food products and ingredients. The initial requirements for an experimentation environment are also introduced.

Keywords: Machine learning · Deep learning · Data analytics · Big data · Experimentation method

1 Introduction

Within the field of artificial intelligence, machine learning algorithms have been widely extended and adopted over the last decade [1]. A driving force for this development have been developments made in the area of neural networks with methodologies such as deep learning [2]. Deep learning networks and their implementations within real-life solutions have a large impact that goes beyond the academic world, giving rise to disruptive approaches in industrial applications as well. However, researchers and developers, who use models and algorithms in a variety of projects and contexts, have yet to find interactive, straightforward and

Published by Springer Nature Switzerland AG 2020
I. N. Athanasiadis et al. (Eds.): ISESS 2020, IFIP AICT 554, pp. 165–172, 2020.
https://doi.org/10.1007/978-3-030-39815-6_16

usable ways in which they may test, execute and parameterise such algorithms, without the need for source code adaptations or additional software installation. Testing methods and tools that may support their systematic implementation and evaluation in the context of near real-life applications are still limited. Experimental testing for these algorithms could be greatly facilitated by interactive experimentation environments that can also offer the computational, memory and storage power that large scale simulations require [3].

In the past, a number of software toolkits and frameworks have been proposed trying to address this need in the area of information filtering systems. Examples include systems like Lenskit [4], MyMediaLite [5], CollaFiS [6]; they are experimentation environments that can be used to set up, parameterize, and evaluate a variety of algorithms over a number of synthetic or real life data sets. In most cases, these have been software libraries and frameworks rather than actual experimentation environments that provide a graphical user interface. They also do not support preparatory tasks such as pre-processing, normalising and splitting the data sets into train and test components. Especially in the context of commercially deployed platforms for data analytics, it is challenging to experiment with a variety of algorithms, to test a variety of parameters to select the ones that seem to perform better over real data, and to customise the version(s) appropriate for each intelligent decision support feature.

In this paper, we propose an extension to the big data platform that supports food data analytics and intelligence. More specifically, we consider the addition of a Prediction Experimentation Panel that has a graphical user interface to help our data scientists, as well as researchers and developers that we collaborate with, test a variety of prediction algorithms in an easy and customisable way and then be able to use a user-friendly dashboard to evaluate experimentation results. First, we describe our existing big data platform for which this extension is considered. Then, we describe a usage scenario that comes from our actual experiments and that is related to food price prediction. Finally, we propose some initial directions towards implementing such a Prediction Experimentation Panel in our big data platform.

2 The Big Data Platform

Agroknow's Big Data Platform is a back-end system responsible for collecting, processing, indexing and publishing heterogeneous food and agriculture data from a large variety of data sources. The platform is organized in a microservice architecture, with different technology components handling different aspects of the data lifecycle. All of the components are interconnected using well-defined connectors and API endpoints, each responsible for storing and processing different types of data. More specifically, the platform includes:

– the **Data APIs** component, which is the machine-readable interface to the different types of data collected in the platform. This part of the architecture is responsible for making data discoverable, but also for submitting new data assets back to the platform,

- the **Data Integration** component, through which data is submitted to the platform through a workflow of four unique steps: data collection, data filtering and transformation, data enrichment and data curation,
- the **Data Indexing** component, which performs data transformation to an appropriate format designed for performance optimization,
- the **Storage** component, which features various storage engine technologies, responsible for the physical archiving of data collections.
- the **Knowledge Classification** component, which provides rules and standards for the organization of data records stored and processed by the platform
- the **Data Processing** component, which is responsible for hosting individual text mining, machine learning and data correlation scripts that can be used in a variety of contexts as standalone pieces of code or as web services through the so-called Intelligence APIs.

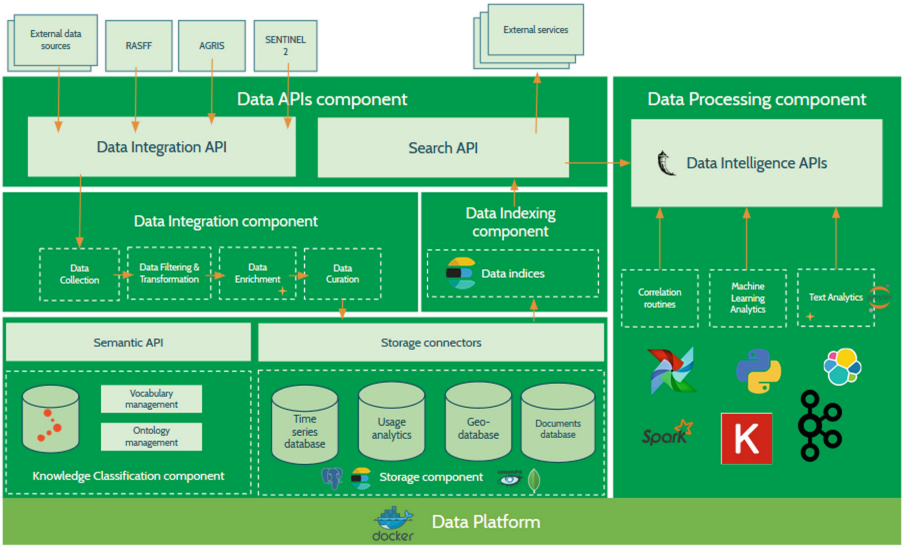

Fig. 1. The Agroknow Data Platform architecture

3 Towards a Method for More Accurate Predictions: The Case of Food Price Prediction

To explore typical design options for parameterization, one can examine the plethora of proposed approaches in machine and deep learning algorithms. These engage various methods and techniques at each step of a prediction workflow,

leading to a wide number of parameters that may be useful for scientific experimentation. Those can also be categorized under two generic steps or processes: the **data preparation** process and the **execution and evaluation** process. Figure 2 schematically illustrates a typical experimentation process that our data science team has to follow in order to run only a handful of experiment iterations for just one of the intelligent features that we examine: price prediction for food products and ingredients. In the sections that follow, we describe some of the required steps in more detail so that we illustrate the complexity of the tasks that we would like to support.

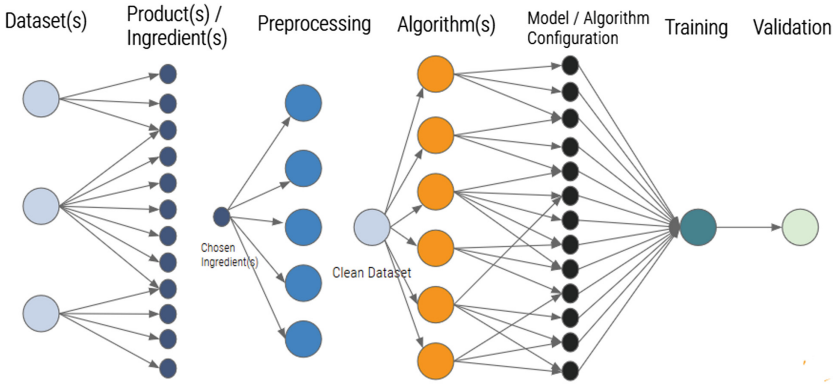

Fig. 2. The workflow followed during our food price prediction experiments

3.1 Data Preparation Process

The data preparation process [7] includes data cleaning, data integration, data selection, and data transformation. *Data Collection* is the process gathering the data that is relevant to the scope of the prediction hypothesis. In the context of the Agroknow Data Platform, data is periodically harvested from different data sources, that may provide it in various formats and mediums. For the food price prediction case, data were collected from the Hellenic Food Market, the European Commission and the Food and Agriculture Organization of the United Nations. *Data collection* corresponds to the *Dataset(s)* in price prediction experiment workflow (Fig. 2). *Data Cleaning* is the process of cleaning, filling the missing values and repurposing data, so that it is well-presented and concise, according to machine-readable standards. Our big data platform performs data cleaning in the way of transforming and filtering values, so that only relevant ones can then end up in machine-readable formats (eg. APIs). An extra process deals with the enrichment of missing or inadequate values of the data. In the case of food price prediction, lots of missing or malformed records and outliers were detected, thus resulting in incomprehensible values, which were automatically corrected. The

finalized product was then curated by data experts that validated before making it available for the final step. *Data cleaning* corresponds to the *Preprocessing* in price prediction experiment workflow (Fig. 2).

Data Transformation and Selection: To bring the data in the appropriate format (that is going to be consumed by the prediction model), data needs to follow a specific structure that is understandable by the prediction model. In modern data platforms, this task can be dealt with the use of data indexing mechanisms. Our big data platform made use of its Data Indexing component for bringing slices of data in the desired format. The food price prediction model requires additional data normalization, that is carried out by the indexing software used with the provision of data aggregations or statistics of the indexed data. *Data Transformation and Selection* corresponds to the *Model/Algorithm configuration* in price prediction experiment workflow (Fig. 2).

3.2 Execution and Evaluation Process

Execution and evaluation process includes data mining, pattern evaluation, and knowledge representation. In **Data Mining process**, we have applied methods to extract patterns from the data. Also, this mining includes several tasks, such as classification, prediction, clustering, time series analysis, and so on. In food price prediction case we develop time series prediction algorithms. In this step have been tested six different algorithms that was Moving Average (Standard vs Exponential), K-Nearest Neighbors, LinearRegression, Arima, Long short-term memory and Prophet for more than 800 products. For each algorithm, the data were transformed and consolidated into different forms that are suitable for mining. All these configurations were done directly in the code. The algorithm was executed by running the algorithm in a python environment *Data Mining process* corresponds to the *Algorithms(s)* and *Training* in price prediction experiment workflow (Fig. 2). Last but not list is the **Evaluation** that validates the algorithm's result. The process of evaluating data using analytical and logical reasoning to examine each component of the data provided. If the results are not as expected then the researcher should go back to the previous steps and make changes. *Evaluation* corresponds to the *Validation* in price prediction experiment workflow (Fig. 2).

3.3 Designing a Prediction Experimentation Panel

The Prediction Experimentation Panel helps researchers and developers to develop the processes between transformation and evaluation process via a user-friendly platform. In particular, the user (researcher or developer) can choose from a drop-down menu the dataset that is clean and ready to use, the algorithm that the user wants to experiment with and then, fill in the parameters that each algorithm needs. Finally, the user executes the algorithm. The execution is part of data processing component and in particular part of machine learning analytics (Fig. 1). When the execution finishes the user can evaluate

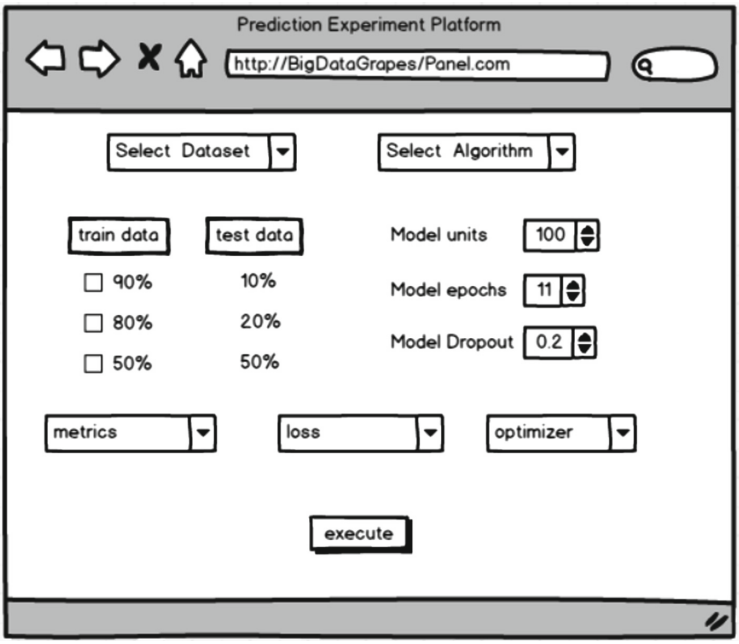

Fig. 3. Initial mockup for prediction experimentation panel

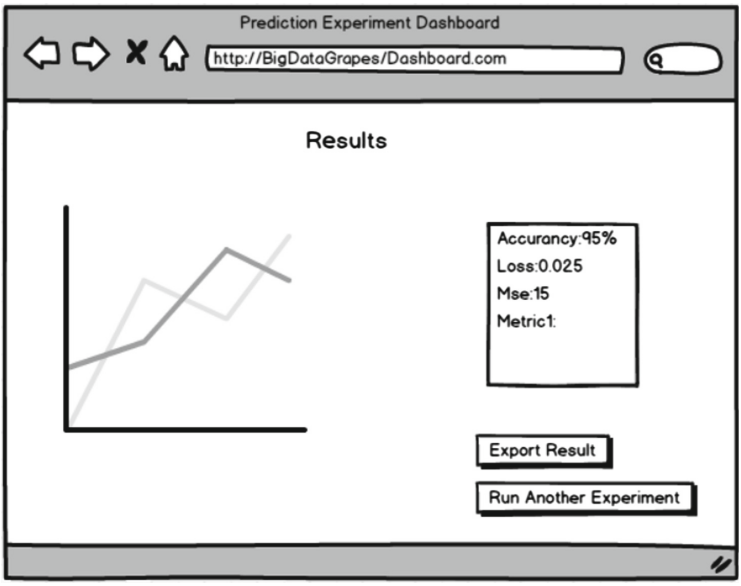

Fig. 4. Initial mockup for prediction experimentation dashboard

the experiment's results using the platform' s dashboards. Results storage in database storage component (Fig. 1). Dashboard includes charts that is line, pie, area, bar, scatter, etc and comments about metrics in order to evaluation the algorithm efficiently.

Data Selection: When a user select a dataset from a drop-down menu, in particular she sends a request to the database through the search API and the response is the selected dataset. The user has the opportunity to parametrize the request by filling in some fields such as date duration and data source. As a result, the response will be a sub-dataset with specific duration and data source. Otherwise, the response is the whole dataset. These parametres are dynamically change in order to fit with the dataset's metadata.

Algorithm Selection: Regarding algorithm selection, user select the algorithm that prefers and fill in the parametres. These parametres dynamically change based on each algorithm needs. This process is part of data processing component (Fig. 1). In a few words the panel adapts dynamically to the requirements of datasets and algorithms Finally, the evaluation involved comparing diagrams of different algorithms.

This process is quite time consuming and developers or researchers need to change the code to make changes to some parameters of an algorithm or a dataset. When it comes to evaluating results, comparing diagrams and metrics is quite difficult when you have so many different algorithms and datasets.

Figures 3 and 4 show initial mockups for prediction experimentation panel. In order to execute Long short-term memory algorithm for price prediction experiment, the panel must have the following parameters which is dataset and algorithm selection from a dropdown menu, the train and the test data, metrics, loss and optimizer in order to evaluate the model (Fig. 3). These parameters have dynamically change for a different algorithms because needs are different. Results are included in a dashboard and the user can export them (Fig. 3).

The figures that are included illustrate the way that we are designing the Experimentation Panel, through a number of graphical mockups that are guiding our implementation work. In order to design a user-friendly Experimentation environment helps researcher and developers to support their experiments, we are in contact with researcher as well as developers that is Agricultural University of Athens (AUA), National Institute of Agricultural Research (INRA), National Research Council (CNR) to get an efficient design for the panel. For this reason, after the implementation of the first demo there will be an evaluation by the users I mentioned above to make the necessary improvements.

4 Conclusions and Next Steps

The implementation of the Experimentation Panel within our big data platform is already undergoing. Our intention is to demonstrate at the workshop a live demonstration of the food price prediction experiment, using the new capabilities of our platform. We believe that such approaches can greatly facilitate the

work of data scientists and software developers in companies like ours, as they offer a useful and practical tool that supports extensive experimentation. As part of our R&D work, we are also incorporating novel machine learning algorithms developed by colleagues in academic environments. We want to offer a joint environment in which we can use the rich and heterogeneous data that our big data platform is continuously collecting and processing, to support the experiments of other colleagues in the community. This Experimentation Panel can be an extension to other similar data platforms.

Acknowledgements. This work is funded with the support by European Commission, and more specifically project Big Data Grapes "Big Data to Enable Global Disruption of the Grapevine-powered industries" (Grant No. 780751) (http://www.bigdatagrapes.eu/), which is funded by the schema "Research and innovation actions (RIA)" under the work programme topic "ICT-16-2017 - Big data PPP: research addressing main technology challenges of the data economy". This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use, which may be made of the information contained therein.

References

1. Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv preprint. arXiv:1708.08296 (2017)
2. Feelders, A., Daniels, H., Holsheimer, M.: Methodological and practical aspects of data mining. Inf. Manag. **37**(5), 271–281 (2000)
3. Manouselis, N., Stoitsis, G.: Towards an e-science environment for collaborative filtering researchers. Int. J. Digit. Libr. Syst. (IJDLS) **4**(1), 41–72 (2014)
4. Ekstrand, M.D., Ludwig, M., Kolb, J., Riedl, J.: LensKit: a modular recommender framework. In: RecSys (2011)
5. Gantner, Z., Rendle, S., Freudenthaler, C., Schmidt-Thieme, L.: MyMediaLite: a free recommender system library. In: RecSys (2011)
6. Manouselis, N., Costopoulou, C.: Designing a web-based testing tool for multicriteria recommender systems. Eng. Lett. Spec. Issue Web Eng. **13**(3) (2006)
7. Kietz, J.U., Serban, F., Bernstein, A., Fischer, S.: Data mining workflow templates for intelligent discovery assistance and auto-experimentation. In: Third-Generation Data Mining: Towards Service-Oriented Knowledge Discovery (SoKD-10), pp. 1–12 (2010)