ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Machine learning for marketing on the KNIME Hub: The development of a live repository for marketing applications

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Villarroel Ordenes, F., Rosaria, S. (2021). Machine learning for marketing on the KNIME Hub: The development of a live repository for marketing applications. JOURNAL OF BUSINESS RESEARCH, 137, 393-410 [10.1016/j.jbusres.2021.08.036].

Availability:
This version is available at: https://hdl.handle.net/11585/981472 since: 2024-09-06

Published:

DOI: http://doi.org/10.1016/j.jbusres.2021.08.036

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

15 September 2026

Marketing Analytics on the KNIME Hub: The Development of a Live Repository for Learning, Sharing, and Reusing Machine Learning Projects for Marketing

Francisco Villarroel Ordenes, Assistant Professor of Marketing, LUISS Guido Carli

Rosaria Silipo, Head of Evangelism, KNIME

(Accepted Version)

Keywords: Machine Learning, Marketing Analytics, Computer Vision, Text Mining, Deep Learning

Acknowledgments: The authors thank Prof. Michael Berthold, CEO of KNIME for his feedback on earlier versions of this manuscript, Roberto Cadili for his work on the Google Cloud Vision workflow, Alì Marvi for help with keyword research for SEO, and Kathrin Melcher for assistance in using deep learning for sentiment analysis.

Machine learning (ML) and data analytics together accounted for the most investments in digital marketing activities in 2020, in line with expected investment increases of 20% in the next three years (Deloitte, 2021). Yet nearly half of chief marketing officers allege that such investments have no real business impact (Deloitte, 2021). According to Mela and Moorman (2018), the reason for such perceptions stems from the lack of organizational capacity to perform the complex tasks related to ML in marketing, a prediction that appears supported by the short supply of professionals trained in such methods (Marketing Tech Advisor, 2019). Data scientists also appear unable to develop or collaborate effectively on marketing projects. Such challenges highlight the need for applied research guides to accelerate learning and collaboration on ML projects for marketing purposes.

We find some prior studies that seek to develop macro frameworks to facilitate conceptualizations of ML in marketing (Ma & Su, 2020; Proserpio et al., 2020), as well as micro analyses of ML methods applied to tackle specific problems, such as customer churn (Amin et al., 2019), online sentiment (Hartmann et al. 2018), or video analytics (Li et al., 2019). Between these two perspectives, we take a median view, seeking to facilitate learning and implementing various marketing-relevant ML tasks in a single environment. Some recent research suggests ways to enable ML tasks using code-free tools (Ciechanowski et al., 2021), but those tools might not be suitable for more complex ML tasks, nor for enhancing collaboration within the marketing community. To address these gaps, we seek to develop a live repository of ML-marketing projects, which can facilitate learning, sharing and reusing projects in a single environment that promotes collaboration. The repository is developed using the KNIME Analytics Platform, an open-source visual programing platform, that can help data scientists (1) learn in an intuitive

way, (2) collaborate, and (3) solve complex ML projects. These efforts provide three novel contributions to ML literature in marketing.

First, we extend previous machine learning literature by conceptualizing visual-based programing that can account for coding interfaces that are not entirely written or script-based (e.g., R, Python) but that instead are visually organized according to their uses of configurable nodes (e.g., to perform regressions) and arrows (e.g., connectors in sequence). Drawing on learning theory (Krätzig & Arbuthnott, 2006) we posit that, to extend and amplify the use and understanding of ML in marketing, users should have alternative learning modes beyond script-based coding. Visual programing underlies the KNIME Analytics Platform, which we use to exemplify the applicability of visual learning in ML projects.

Second, we advance previous marketing research concerning ML macro conceptualizations, and specific tutorials, by developing a live repository for researchers and practitioners to learn, share and reuse projects at the intersection of marketing and ML. To develop this repository, we leverage Ma and Sun's (2020) ML framework that addresses different marketing needs and methods concerning supervised, unsupervised and deep learning, among others. The repository is called "Machine Learning and Marketing" and it includes a selection of five curated projects, including one or more workflows (i.e., visual scripts) in the marketing areas of customer churn, sentiment analysis, automated analysis of images, SEO and customer experience. We call it a live repository, because it is designed to grow overtime with the inclusion of new workflows related to ML and marketing. The projects are hosted within the KNIME Hub, which is an active, collaborative, global community of data scientists. As its main advantages, this repository (1) requires little code writing, which hastens the learning cycle for academics and students; (2) allows the implementation of a wide range of ML tasks across

different types of data; and (3) supports rapid integration with other environments within KNIME, such as R, Python, Amazon AWS, or Google Cloud.

Third, we developed two detailed step-by-step guides (i.e., cases) describing the customer churn project, and one of the sentiment analyses projects that relies on text mining and supervised machine learning. These step-by-step guides contribute to fasten the learning curve of new KNIME Analytics users and enthusiasts in ML. With these step-by-step guides, users can identify the basic steps for executing their own ML projects in KNIME. For each case, we introduce key marketing issues and link them with pertinent marketing research, then we describe the data, and provide a detailed explanation of the steps required to build and deploy a workflow, to guide their usage by academics, managers, or students.

In the next section, we conceptualize visual coding and introduce KNIME Analytics Platform. Using the framework proposed by Ma and Sun (2020), we then introduce the KNIME Hub and a list of ML marketing tasks available for research projects. We continue by presenting the step-by-step development of two ML cases for marketing, pertaining to customer churn and sentiment analysis. Finally, we conclude with brief descriptions of the other relevant ML projects in the hub and some final thoughts for its development in the near future.

1. Visual Programming for Marketing Analytics

1.1. Conceptualizing the visual coding

Learning theories suggest that people have preferred modalities for absorbing, retaining, and processing information (Krätzig & Arbuthnott, 2006). These modalities can be classified into seven styles: visual (images), logical (mathematical), verbal (linguistic), physical (kinesthetic), aural (auditory), social (interpersonal), and solitary (intrapersonal) (Davis, S. E. 2007). The

styles in turn can work in combination, depending on the person's preferences or abilities; furthermore, there is no evidence of the superiority of one learning style over another (Krätzig & Arbuthnott, 2006). Although real-world practices of ML and data science tend to be dominated by written and logical modalities (i.e., coding scripts), a new generation of platforms that feature the intersection of visual and logical modalities have gained popularity, such as KNIME Analytics Platform, RapidMiner, and AzurreML Studio. For this study, we develop the empirical research and cases on the KNIME Analytics Platform, which uses a visual environment to support data science projects. The KNIME Analytics Platform combines together four important features that are relevant for this project: (1) it is open source and free, (2) it offers a visual programming, (2) it relies on contributions by the community through an open and free repository (the KNIME Hub), and (4) it is built on an open architecture that allows for seamless integration with external tools and data sources (e.g., R and Python Script). Table 1 provides a comparative analysis with other platforms to perform analytics. We thus delineate visual learning properties that may be particularly relevant when applying ML to marketing projects.

[Insert Here Table 1]

Human brains are mainly image processors (most of the sensory cortex is devoted to vision; MIT News, 1996), so the presence of visual objects can hasten content comprehension (Hattie, 2011). Several studies confirm the positive effects of visual imagery on learning outcomes; for example, Yen et al. (2012) demonstrate that image-based (cf. text-based) learning results in greater understanding and creativity in concept mapping tasks. Because ML is a fairly complex subject, characterized by a broad range of programing languages (e.g., JavaScript, Python), data types (e.g., numbers, text, images, audio), methods (e.g., supervised, unsupervised, mixed), and tools (e.g., AWS, Scraping API's), we argue that visual learning might simplify ML

implementations in business domains. A visual interface saves programing time (e.g., script writing), which instead can be used to comprehend the business problem more clearly or learn about different algorithm alternatives. This interface does not eliminate mathematics, logic, or statistical knowledge, but it packages them within visual *nodes* and configuration options, for a more streamlined learning experience.

Visual programming also can facilitate collaboration among members of a team, despite their different skills or backgrounds. If the work is highly collaborative and integrative, a visual programming environment allows the team to model and document all their work in a single, consistent way. Visual workflows also can abstract away from different, underlying types of code (e.g., SQL, Python, R, JavaScript) or environments (e.g., AWS, Google Cloud), to focus on creating a process for (1) accessing data, (2) organizing and aggregating data, (3) extracting insights and models, and (4) putting the results into practice or production. That is, visual programming for data science achieves an appropriate level of abstraction but maintains the complexity required by data scientists. It also implies that various people involved in creating a data science process can collaborate and share their specific expertise, while still working in a single, common, consistent environment.

1.2. KNIME Analytics Platform

KNIME Analytics Platform is based on five main pillars (Berthold, 2014; Berthold et al., 2020): an open architecture, open-source philosophy, extensions of algorithms, extensions of operations covered, and (most pertinent to our research) visual programming, which supports its transparency, collaboration, agility, and power. This open-source platform offers all required

functionalities to take data, transform and analyze them, then produce relevant results.[1] As the KNIME workbench in Figure 1 shows, to create an empty canvas for a new project, users right-click on LOCAL in the KNIME Explorer panel or else click on File → New in the top menu, then select "New KNIME Workflow."

[Insert Figure 1]

The basic processing unit in KNIME Analytics Platform is the node. Each node implements a specific task, such as aggregation, row selection, or training a neural network. A node can be created, via drag and drop, from the Node Repository panel into the Workflow Editor panel or by double-clicking in that Node Repository panel. In turn, each node has three possible states, represented by a traffic light displayed under each node (Figure 2): Red indicates not configured, yellow means configured but not executed, and green is successfully executed. A fourth status can appear if the execution includes an error. That is, after creating a node, it must be configured through settings. A setting can be the path of the file to read, a parameter in a ML algorithm, filtering criteria to extract specific rows from the data set, and so on. To configure a node in the Workflow Editor, users can double-click or right-click it and select "Configure," then insert the configuration parameters. If configuration is successful, the traffic light changes to yellow; the configured node is ready, but it has not yet performed its task. Finally, the user executes the node by right-clicking and selecting "Execute" or by pressing one of the green buttons in the tool bar at the top. If execution is successful, the traffic light changes to green.

[1] Available at https://www.knime.com/downloads.

[Insert Figure 2]

A workflow usually starts with a node that imports data (e.g., file reader, database connector, GET Request to a web service, cloud connector). Then a second node can be created, using the same process, but the output of the previous node gets connected as input to this second node. Specifically, the user clicks on the output port of the first node and releases on the input port of the second node, which establishes a data flow through this connection. The second node still must be configured appropriately and executed, as done for the first node.

Node after node, a pipeline of requested actions thus develops, which produces a workflow from data to the final output. After each execution, users can inspect the results of each node by right-clicking it and selecting the last option in the context menu, such that they can monitor and address any flaws at each step in the process, through the workflow. A good practice also includes comments about the task assigned to each node or group of nodes. To create such an annotation, users can right-click anywhere in the Workflow Editor and select "New Workflow Annotation," then apply their preferred border, fonts, or background with the editor tool.

Finally, KNIME Analytics Platform includes metanodes and components[2] that comprise other nodes. They can isolate logical, self-contained, reusable parts of different workflows. For example, parts that calculate standard customer key performance indicators (KPIs) might be established as metanodes for gathering customer data as input and producing the KPIs as output. Components can have configurable settings, such as the number of KPIs to produce, but they effectively camouflage the complexity of KPI calculations and allow others to reuse these components with little effort if they require the same KPIs. To create a component, users select

[2] Both metanodes and components help to clean up and organize workflow, yet components have several features that metanodes do not have as indicated in this link: https://www.knime.com/blog/metanode-or-component

all nodes involved in the component's task and encapsulate them (right-click → "Create Component"). The selected group of nodes then constitutes the new component (Figure 3). A double Ctrl click opens the content of the component.[3]

[Insert Figure 3]

With these elements, KNIME Analytics Platform offers hundreds of alternatives for executing tasks related to ML (e.g., autoencoders, word embeddings, text mining, image processing). In Figure 4, we match the ML functionalities of KNIME with Ma and Sun's (2020) ML framework, which reveals that KNIME supports efforts to process different types of data (numbers, text, images, audio) and implement a range of ML learning tasks that are relevant to marketing research.

[Insert Figure 4]

1.3. Workflow repository for KNIME community: the KNIME Hub

Another asset provided by KNIME Analytics Platform is its community. The KNIME Forum allows users to ask questions or seek clarifications and general help. The KNIME Hub instead is the site for sharing documents, workflows, nodes, and components, which have been provided by other members of the KNIME community. In this "code" repository, KNIME users can search for solutions to current data science problems. Currently, the KNIME Hub includes nearly 6,000 reusable workflows related to various ML and other data science tasks.

For this article, we undertook two steps to develop a ML repository for marketing. First, we identified workflows in the Hub related to ML and marketing. Through this content curation

[3] For more information on how to create and execute a full workflow of nodes, interested readers can access the resources available in the KNIME LEARNING page.

step for workflows at the intersection of marketing and ML, we ensure their functionality and reliability for this project. Second, we found specific areas in which a new or updated workflow appeared necessary, and we created a customized ML project for these needs. Therefore, we can establish a more complete set of state-of-the-art ML projects that can be used for research or teaching in marketing contexts. Notably, the selection of ML projects in the KNIME Hub represents the starting point of a wider repository of ML projects in the area of Marketing. We expect to continue developing projects, and motivating the broader KNIME community, to share their work in this hub.

In Table 2, we list the author(s), short descriptions, and links to the workflow for each solution. Users can find them on the KNIME Hub, download them to their own KNIME Analytics Platform, and customize them. The projects are in five relevant marketing areas in which ML implementations are well-known for adding business insights: Customer Churn (Ascarza et al., 2018), Consumer Sentiment Analysis (Heitmann et al. 2020), Automated Analysis of Visuals (Villarroel Ordenes and Zhang 2019; Nanne et al. 2020), Search Engine Optimization (SEO) (Schweidel, Reisenbichler and Reutterer 2021) and Customer Experience (Holmund et al. 2020). In each of these projects, we have included one or more annotated workflows to help users to accelerate their learning curve. For example, the Sentiment Analysis includes four different approaches to execute sentiment analysis projects (e.g., dictionaries, traditional ML, Deep Learning, etc.); and each of these approaches includes one workflow for building a sentiment analysis classifier and another one to put into production.

With the purpose of getting readers familiar with the use of these workflows, in the next section we provide two step-by-step guides; one for the customer churn project (case 1) and one for the sentiment analyses project with traditional ML (case 2). We decided to develop the step-

by-step guides for these two projects because they are consolidated topics in today's ML landscape in marketing (churn, Du et al. 2021; sentiment, Heitmann et al. 2020), and the methods that we cover are still relevant in recent marketing research (random forest, Peng, Cui, Chung, and Zheng, 2020; support vector machine, Homburg, Theel and Hohenberg 2020). The two step-by-step guides will help incoming KNIME Analytics users and marketing analytics enthusiasts to have a better first experience with the platform and the ML topics. We used familiar settings in our effort to help readers gain familiarity with the basic steps required to build a visual workflow, so then they can have a better transition to using other workflows in Table 2.

[Insert here Table 2]

2. ML Cases for Marketing

2.1. Case 1: Customer churn

Extensive marketing research seeks better predictions of customer churn, which occurs when customers stop transacting with the firm (Ascarza et al., 2018). New ML algorithms (Amin et al., 2019), data (de Haan & Menichelli, 2019), and services (Dechant et al., 2019) have reinvigorated research interest in this topic. Using existing customer data (e.g., transactional, psychographic, attitudinal), predictive churn models aim to classify customers who have churned or remained, as well as predict the possibility that new customers might churn, in an automated process. If churn probability is very high and the customer is valuable, the firm wants to undertake actions to prevent this churn, or retention management (Askarza et al., 2018). Therefore, we propose building a simple ML classifier that can distinguish customers who have churned and customers who have stayed.

2.1.1. Customer data. Customer data might include demographics (e.g., age, gender), revenues (e.g., sales volume), perceptions (e.g., brand liking), and behaviors (e.g., purchase frequency). Yet the definition of such predictive variables is not always straightforward, because it depends on the business case. For example, customer churn might be defined differently for a subscription-based business (e.g., Netflix) than for a physical retailer (Walmart), so the set of churn predictors, and thus the necessary customer data, would differ too (Ascarza et al. 2020).

For this example, we rely on a popular simulated telecom customer data set, used in previous research to compare alternative churn models (Amin et al. 2019) and available at https://www.kaggle.com/becksddf/churn-in-telecoms-dataset.[4] In our effort to provide a broader overview of KNIME functionality, we split the data set into a CSV file, which contains operational data such as the number of calls, minutes spent on the phone, and relative charges, and then an Excel file that lists the contract characteristics and churn flag (i.e., if the contract was terminated). Each customer can be identified in the data by area code and phone number.

2.1.2. Practical implementation in KNIME. The workflow in Figure 5 provides a visual representation of the series of steps required to create a churn prediction model with supervised ML, which we detail in this section.

[Please Insert Figure 5]

1. Reading the data. Files with data can be dragged and dropped into the KNIME workflow. In this case, we use XLS and CSV files, but KNIME supports almost any kind of file

[4] Other customer churn data sets used in prior marketing research (Amin 2019) are available at https://www.kaggle.com/abhinav89/telecom-customer or https://www.kaggle.com/blastchar/telco-customer-churn. The first link contains 100,000 observations, so it likely requires more processing time to train the customer churn model. The second link provides a 7,043-observation data set.

(e.g., parquet, json.). The *File Reader* node might help troubleshoot any problems or uncommon extensions for files.

2. Data manipulation and preparation. The *Joiner* node matches two tables using one or more columns shared by both tables (i.e., keys). Users can specify how they want to join the data (inner, right, left, full). With the *Number to String* node, users can convert the "Churn" column into a string, to meet the data type required by the classification algorithm (in this case, nominal). Note that KNIME offers a series of nodes to manipulate data (e.g., string to date or vice versa). At this stage, it is important to obtain descriptive statistics using the *Data Explorer* node (or else a *Statistics* node). In KNIME Analytics Platform, some graphical nodes are dedicated to the creation of plots, charts, tables, and other graphical items. As output, other than data, these nodes produce an interactive view (right-click and select "Interactive View"). The *Data Explorer* and *Statistics* nodes belong to this group of graphical nodes. For example, opening the interactive view of the *Data Explorer* reveals that the churn sample is unbalanced, and most observations pertain to non-churning customers (over 85%)—a common problem in customer churn models (Zhu et al. 2017) that can result in substantial misclassification of the minority class (churner customers). We address the imbalance problem in the next step. Finally, we use a *Color Manager* node, which mainly serves an aesthetic purpose, by allowing users to assign colors to specific rows, based on some criterion. In our case, churners are in red and non-churners are in blue.
3. Training, testing and cross-validation. Supervised ML often includes cross-validation (Ma & Sun, 2020) to avoid overfitting, which might result from the specific characteristics of the training and testing (i.e., holdout) samples. To perform cross-

validation, we would use the *X-Partitioner* node, which splits the data into some desired number of training and testing samples, and the *X-Aggregator* node, which collects the results of each iteration. In this case, we used a cross-validation value of five (in the node configuration of *X-partitioner*), so the data set gets divided into five subsamples. In each iteration, four parts are used for training (80% of the data), and one part is used for testing (20% of the data).[5] In addition, for each training set, we first use the *SMOTE* node, which addresses the problem of the less numerous minority class (churners). The SMOTE node oversamples this minority class by creating synthetic minority class examples (Chawla et al., 2002). Finally, for the training and testing algorithms, we use a *Random Forest Learner* node, which produces a trained model (grey square at the end of the node), then a *Random Forest Predictor* node, which relies on the trained model to predict patterns in the testing data. Random forest is a popular ensemble method in marketing; each individual decision tree is built from a bootstrap of the original data, and then a final prediction results from averaging all the decision trees (Ma & Sun, 2020). We apply a random forest model with 100 decision trees; though the data set for this project arguably is too small for 100 decision trees, we use them mainly for illustrative purposes. Users can try various other algorithms too (e.g., *Decision Tree Learner and Predictor*).

4. The last part of the workflow measures the performance of the trained random forest on all resulting predictions. The *X-aggregator* node collects all predictions and produces the following additional columns: probability of churn, probability of no churn, and prediction (based on the highest probability). The *Scorer* node matches the random forest

[5] In this example, we used a fivefold cross-validation, so the process iterated five times, with different training and testing samples. The output are predictions for each of the five testing samples. In this case, each testing sample was 20% of the entire data set (N = 3,333), so we obtain a total of 3,333 predictions.

predictions with the original churn values from the data set and assesses model quality using evaluation metrics such as accuracy, precision, recall, the F-measure, and Cohen's Kappa (Amin et al. 2019). All these metrics range from 0 to 1; higher values indicate better models. Another metric, the receiving operating characteristics (ROC) curve and its corresponding area under the curve (AuC), are widely used to represent the prediction accuracy of binary outcomes (Netzer et al., 2019). For this example, we obtained a model with 93.8% overall accuracy. The AuC for the corresponding ROC curve is 0.89. These results reflect the default settings of the *Random Forest Learner* node; better predictions might be achieved by fine-tuning the settings in the *Random Forest Learner*. Figure 6 provides the different views of the *Scorer* and *ROC* nodes.

[Insert here Figure 6]

Once a researcher is satisfied with the predictive accuracy of a model, it should be applied to new data for actual churn prediction. Figure 7 includes a brief workflow that demonstrates the deployment of a predictive model on new data. The previously best trained model, which in this case turned out to be the one from the last cross-validation iteration, is read (*Model Reader* node), and data from new customers are acquired. A *Random Forest Predictor* node applies the trained model to the new data and produces the probability of churning and the final churn prediction for the input customers. The workflow concludes with a composite view, produced with the "Churn Visualization" component node. If components contain graphical nodes, they inherit these interactive views and combine them into a composite view (right-click "Interactive View"). The composite view of the "Churn Visualization" node shows predictions for five new customers: Four will not churn (blue), and one will (orange). All the items in this view are connected, so selecting a tile prompts a selection of the corresponding bar in the chart.

More complex interactions and dependencies can be inserted in the composite view of a component node, using appropriate nodes.[6] More details are available in existing KNIME documentation (e.g., KNIME Components Guide, 2020).

[Insert Here Figure 7]

2.1.3. Discussion and further ideas related to ML and churn. The customer churn case can be extended, depending on users' needs, and it also applies to other binary classification problems. A common extension might involve employee attrition and efforts to identify the likelihood of workers leaving the organization. A solution, which relies on visualization and oversampling of the original data (Guirao, 2020), posted on the KNIME Hub consists of two workflows: "Training a Churn Predictor" (https://kni.me/w/GCI4B-AzIczhIwjG) and "Deploying a Churn Predictor" (https://kni.me/w/m9lFUDUeLyLFjpMr). Another extension of the customer churn case might include additional predictors, such as the sentiment transcribed in customer conversations with the firm (de Haan & Menichelli, 2019). Notably, in the next case, we present a use of ML for automated classifications of social media chatter to gauge sentiment valence.

2.2. Case 2: Sentiment analysis

The second case, focused on predicting sentiment valence (positive, negative, or neutral) from unstructured text data, relies on a general process that also can be extrapolated to other classification problems, such as specifying types of social media content (Lee et al., 2018; Villarroel Ordenes et al., 2019), sales influence tactics (Singh et al., 2020), or loan default probabilities (Netzer et al., 2019). Sentiment analysis is a highly relevant empirical application for text mining and natural language processing efforts (Feldman, 2013). In marketing, it mainly

[6] Please refer to the following documentation for more information about components: https://docs.knime.com/2020-07/analytics_platform_components_guide/index.html

has been used to measure consumers' expressed sentiment about products, services, or brands (Villarroel Ordenes et al., 2017), as contained in online data such as reviews and social media conversations through Twitter or Facebook.

In reviewing previous sentiment analysis applications in marketing research, we broadly distinguish two different implementations: (1) *measuring sentiment intensity* by computing the proportion of negative or positive words in a document (e.g., using dictionaries such as LIWC [Pennebaker et al., 2015], evaluative lexicon [Rockagle et al., 2018], or VADER [Herhausen et al., 2020]) and (2) *predicting sentiment valence* (e.g., positive vs. negative; positive vs. negative vs. neutral) using different forms of ML.[7] Both implementations can be conducted within KNIME, but we focus on the second, because sentiment intensity is relatively easily computed using paid (LIWC) or open-source (Evaluative Lexicon) dictionaries, without requiring analytics, and because a sentiment analysis application with such dictionaries is already available on the KNIME Hub, as indicated in Table 2.

Recent research by Heitmann et al. (2020) distinguishes four approaches for predicting sentiment valence: (1) lexicons (using dictionaries and rules), (2) traditional ML methods (e.g., support vector machines [SVM]), (3) artificial neural networks, and (4) language models (e.g., Google BERT). We add sentiment prediction by ensemble learning methods, such that users can combine several ML algorithms (Lee et al., 2018). Their accuracy then depends not just on the algorithm but also on contextual factors, such as the text source (e.g., tweets vs. reviews), type of preprocessing (e.g., word stemming), text length, language, and number of training samples (Heitmann et al., 2020). We develop and curate workflows using each approach (Table 1); in the

[7] Sentiment predictions also are possible by using the rules in sentiment dictionaries (Heitmann et al., 2020)

following description, we focus on traditional ML, and we also recommend starting with this method and then moving to other approaches.

2.2.1. Customer data. The sentiment analysis project requires two data sets. One annotated data set that trains the predictive model; it should contain at least one column with text data and another with the sentiment annotation (e.g., positive, negative, neutral). The second data set should not be an annotated one, which is necessary for deployment (i.e., to predict the sentiment from the text column). For this example, we use a popular annotated data set used in previous marketing research to compare alternative sentiment analysis models (Heitmann et al., 2020). The data include customer tweets about six major U.S. airlines, scraped in February 2015; they are available at the following Kaggle link: https://www.kaggle.com/crowdflower/twitter-airline-sentiment. Each tweet was annotated by Kaggle contributors as positive, negative, or neutral. Another column indicates the contributors' reasons for choosing that classification. The non-annotated data set is scraped using the Twitter API nodes, within KNIME.[8]

2.2.2. Practical implementation in KNIME. The workflow in Figure 8, Panel a, provides a visual representation of the series of steps required to create sentiment analysis models using supervised ML. Hereafter, we describe each of the steps annotated in the workflow.

- Reading the data. As in the churn example, the data file can be dragged and dropped. In this case, it is a single CSV file that consists of 14,640 tweets.
- Data manipulation and preparation. Duplicate rows are a common problem when scraping social media data, so the *Duplicate Row Filter* node functions to identify and exclude such repetitions. The *Strings to Document* node converts a series of string

[8] Users can download the "Twitter API Extension" in Knime (File → Install KNIME Extensions). Users need a Twitter API (with credentials) through the Twitter developer's page https://developer.twitter.com/en/apply-for-access

columns (e.g., tweet text, tweet sentiment, author) into a single document column, which is the required format for most text mining tasks in KNIME (Tursi & Silipo, 2017).

- Enrichment and preprocessing. Using enrichment, users can tag words with part of speech (e.g., nouns; *POS Tagger* node), names and entities (e.g., organizations; *Open NLP NE Tagger* node), n-grams (e.g., negations such as "not good"; *NGram Creator* node), and any desired word dictionary (*Dictionary Tagger, Wildcard Tagger* nodes). In this case, we tagged positive and negative words obtained from the MPQA subjectivity lexicon (Wilson et al., 2005), to ensure sentiment-laden words are not removed during the preprocessing step.[9] The preprocessing includes several commonly used operations (Berger et al., 2020), such as removing punctuation, stop words, and words with fewer than four characters. In addition, we include a metanode with a process to remove rare or infrequent words. The threshold for infrequent words can vary, but it is important doing it to fasten the execution of ML algorithms. We use these options as an example, but users can engage in more or less aggressive cleaning, depending on the characteristics of the project (e.g., projects with extreme amounts of data might benefit from word stemming or lemmatization).
- Bag of words and document vectors. After preprocessing text data, the *Bag of Words Creator* node operates on the "preprocessed" document to create a long table with all words or terms in the data, each one in a single row. We then convert terms, into strings and use the *TF* node to compute term frequencies within a document (which can be configured as an integer or a weighted value, relative to the total number of terms).[10]

[9] Some stop word lists include words such as "good" or "bad," which are clearly valenced and might be necessary to predict valence.

[10] Users can choose other measures, such as TF-IDF to assign more weight to infrequent terms (Berger et al., 2020).

Finally, the *Document Vector* node transforms the bag of words into a table; each row is a single document, and the word/terms appear in columns. The representation of documents as word vectors helps in using all or some of these vectors as predictors by the ML algorithm. The conclusion of this metanode entails outputting a Document Vector table, which represents the "model" of the document vector space to apply to the deployment workflow for non-annotated data.

- Supervised machine learning and evaluation. The last metanode, *Supervised Machine Learning (Save models)*, trains an SVM algorithm on 80% of the documents, then tests it on the remaining 20%. To simplify the workflow, we did not apply *cross-validation*, but it is possible, similar to the customer churn case. In addition, a parameter optimization loop might be applied to select the best parameters for different types of SVM models (linear or nonlinear) available in the *SVM Learner* node configuration. We check model accuracy, which reaches 78.3% in this case, in line with previous marketing studies that address this classification problem (Heitmann et al., 2020).

With the sentiment model, users can apply their predictor to new, non-annotated data. Figure 9 includes a workflow that demonstrates how to deploy a predictive model to estimate the sentiment of recent tweets scraped with the Twitter API extension nodes from KNIME. It starts with the component *Tweet Extraction*, including the node *Twitter API Connector* to connect to the Twitter API and the node *Twitter Search* to query the API for tweets with a given hashtag[11]. The component has been implemented to be configurable. Configuration nodes within the component create the configuration dialogue of the components for the Twitter credentials and

[11] Users need a Twitter API (with credentials) through the Twitter developer's page https://developer.twitter.com/en/apply-for-access

the search query. By default, the Twitter API returns the tweets from last week, along with data about the tweet, the author, the time of tweeting, the author's profile image or number of followers, and the tweet ID. For each tweet, a document gets created in the *Strings to Document* node, which features all tweet information and the author profile image. Next, the metanode *Enrichment and Preprocessing* follows the same steps as the metanode with the same name that we described in the workflow that trains the model. The metanode *BoW and Vector Space* then extracts the bag of words from the whole tweet set and calculates the absolute frequencies of each word within each tweet. Given the fixed (short) length of tweets, relative frequency offers no additional normalization advantage for the frequency calculation. With the corresponding Document Vector, SVM can be applied, using the relevant models created in the training workflow. Finally, the *Document Data Extractor* node retrieves all tweet information stored in the document, joins the profile image back to the tweet, and allows the workflow to continue with data visualization. The *Visualization* component produces a word cloud of the 150 most frequent terms in the tweet corpus; a bar chart with the number of negative, positive, and neutral tweets; and a table with all extracted tweets color coded by predicted sentiment (Figure 10).

2.2.3. Discussion and further ideas for ML for unstructured data. The workflows developed for sentiment analysis can be extrapolated to similar automatic classification problems for textual data. For example, classifications of social media brand messages into content types (Villarroel Ordenes et al., 2019; Lee et al., 2018) or selling tactics (Singh et al., 2020) would require similar processes, which could be implemented using any approaches for sentiment classification. Further research could go beyond analyses of single sentences and messages, to classify dialogue turns or entire conversations between customers and employees. This type of analysis, involving pairs of human interactions, would be useful for developing more complex constructs, such as linguistic empathy.

2.3. Other ML marketing applications in the KNIME Hub

With the two cases, we sought to introduce the use of KNIME for ML. In addition to these detailed examples, we have curated several workflows to help researchers and practitioners implement other ML tasks related to marketing problems. We select three relevant marketing areas and curated workflows designed to tackle the problems using ML: customer experience using topic models, social media content marketing using image mining, and SEO using Google search data and topic models.

2.3.1. Customer experience and topic models. Customer experience management and customer journeys are highly relevant marketing topics in recent years (McKinsey 2021), and various articles seek to conceptualize the underlying notions (Homburg et al., 2017; Siebert et al., 2020; Verhoef & Lemon, 2016) and how to measure them using big data analytics (BDA) (Holmund et al., 2020). Topic models (Blei, 2012) consistently are applied in research to learn about customer experiences, though most implementations rely in online reviews (Sutherland &

Kiatkawsin, 2020; Timoshenko & Hauser, 2018) or open-ended responses to questionnaires (Piris & Gay, 2021; Villarroel Ordenes et al., 2014). Their popularity in turn has prompted continuous developments of algorithms such as latent Dirichlet allocation (LDA), correlated topic models (CTM), and structural topic models (STM) (e.g., Büschken, Allenby 2020), which also have been implemented in business research. The multitude of topic model methods and increasing interest in measuring customer experience motivated the development of a KNIME workflow for customer experience that allows users to visualize the implementation of LDA, as well as the R integration, within KNIME. This workflow, in Figure 11, depicts the different steps involved in implementing a topic model and supports comparisons of the results of two different implementations, using evaluation measurers (i.e., perplexity) and visualizations. Panels a and b contain the workflow and then the results.

[Insert Figure 11 Here]

2.3.2. Content marketing and image mining. We note exponential growth in the amount of visual data available, including images and videos. In turn, new technologies aim to classify and extract relevant insight from images; the image recognition market (largely driven by ML) is expected to reach a value of US$5,161 million by 2026, representing 24.82% growth between 2021 and 2026 (Mordor Intelligence, 2020). Because consumers and firms rely more on pictorial information and videos to communicate (Grewal et al. 2021; Villarroel Ordenes and Zhang 2019), researchers need new processes and methods to analyze these data, including information contained in consumer selfies (Hartmann et al., 2020), customer service complaints (Li & Xie, 2020), promotional pictures for shared services like Airbnb (Zhang et al., 2020), or firm-generated content in social media (Villarroel Ordenes et al., 2019). Interest in being able to

analyze visuals and their implications for firm performance prompted the creation of a workflow for analyzing visual content. It takes advantage of Google Cloud Vision services[12], which rely on ML to detect labels (e.g., humans), and extracts nuanced image properties such as color concentration. Identifying image objects/subjects and color properties (Labrecque et al., 2013; Li & Xie, 2020) remains a prominent goal for marketing research, and the workflows can support more empirical work in this area. Figure 12 includes the Google Cloud API workflow and the results it produces; Table 1 also cites other workflows for image classification tasks that use deep learning and pretrained models, which could be adapted to perform image classification tasks in marketing content (Hartmann et al., 2020; Liu et al., 2020).

[Insert Figure 12]

2.3.3. Keyword research for SEO. Search engines rank web pages and other brand content (e.g., videos; Cowley, 2019), according to the presence of specific keywords or groups of keywords that are conceptually and/or semantically related. Marketers should regularly seek the best keyword ideas, to validate that consumers can find them and determine their competitiveness. Keyword searches tend to be guided by online explorations of concepts related to a focal service, product, or content (Hubspot, 2020). Popular sources for keywords are SERP (Search Engine Result Pages) and social media networks such as Twitter and Reddit. On the KNIME Hub, users can find a workflow (Table 1) for semantic keyword research, which resulted from a recognition of the lack of marketing analytics content related to SEO. It can suggest potential topics and wordlists for SEO. The upper branch of the workflow (Figure 13, Panel a) links to Twitter and extracts the most recent tweets that contain a selected hashtag (e.g.,

[12] Users need a google vision API and download its JSON file with the service account key https://cloud.google.com/vision/docs/setup

#cybercrime). The lower branch connects to the Google Analytics API and extracts SERPs around a given search term (cybercrime). The subsequent data flow is the same for both branches: URLs are isolated, web pages scraped, and keywords are extracted together with their frequencies. The component *Retrieve Text* uses a list of URLs as input, sends it to Boilerpipe API via GET Requests for web scraping, and returns the pure text contained in the web pages.

From these texts, the component *Keyword Research* extracts keywords that represent: (1) single terms with the highest TF-IDF (term frequency * inverse document frequency) score; (2) co-occurring terms with the highest co-occurring frequency; (3) and keywords with the highest score from topics detected using the LDA algorithm. The *Keyword Visualization* component then depicts the top single, co-occurring, or topic-related keywords, as shown in the word cloud in Figure 13, Panel b. These results identify which words to include on a web page that relates to cybercrime services to increase the firm's page ranking.

[Insert Figure 13]

3. Limitations, further research and other marketing analytics applications in the KNIME Hub

Our research is not exempt of limitations. While the focus of our research is the development of a living repository where marketing researchers and practitioners can learn, share and reuse workflows (together with two step-by-step applications), new KNIME users might still need to complement these materials with freely available KNIME tutorials (e.g., KNIME YouTube channel) and books. As with any programing language, KNIME offers several free books and course material for researchers and academics. Users can refer to this material in

the KNIME Academic Alliance[13]. Furthermore, this article is not exhaustive of all the topics that can be covered in ML (Ma and Sun 2020). As such, we see our work as a first step to foster sharing and collaboration among researchers interested in marketing and ML. We hope that with time this repository will be populated with workflows that researchers and lecturers can reuse in their day-to-day work. Finally, KNIME Analytics, as any programing language, has limitations too. According verified Gartner reviews[14], KNIME Analytics platform has scope to improve in the following areas: Optimization, Data Exploration and Visualization, and Pre-canned Solutions. However, this is where the KNIME integrations such as Python and R might be very helpful. For example, it is well-known that R is a great source of modelling and statistics. Despite KNIME doesn't offer specific modelling resources (e.g., negative binomial regression or tutorials to handle endogeneity), these types of tasks can be performed in KNIME by using the R nodes integrated in KNIME.

Because ML is a constantly evolving field, researchers interested in state-of-the-art implementations, such as autoencoder models (e.g., ProLDA), word embedding (e.g., Word2Vec), and network embedding (e.g., Spectral Clustering) (Ma & Su, 2020) can benefit from recent implementations available in the KNIME Hub.[15] These models and algorithms are still very new to marketing, with limited implementations in prior literature (e.g., word embedding, Timoshenko & Hauser, 2019; autoencoder, Verboven et al., 2020). We did not

[13] https://www.knime.com/academic-alliance
[14] https://www.gartner.com/reviews/market/data-science-machine-learning-platforms/vendor/knime/product/knime-analytics-platform/ratings
[15] An implementation of an autoencoder for fraud detection is available at https://hub.knime.com/knime/spaces/Examples/latest/50_Applications/39_Fraud_Detection/03_Keras_Autoencoder_for_Fraud_Detection_Training~9qFNMrsuN4PH1hRg. An implementation of word embedding for sentiment analysis is available at https://hub.knime.com/knime/spaces/Examples/latest/04_Analytics/14_Deep_Learning/01_DL4J/08_Sentiment_Classification_Using_Word_Vectors~HYlUnc7CnzG2fe8d

curate, nor developed KNIME cases for these models, but we encourage researchers to develop implementations of the state-of-the-art models and share them in the KNIME Hub.

Our focus is on marketing applications involving ML; other marketing analytics topics (not necessarily using ML) are worthy of study too. For example, a relevant topic that has received little attention in marketing analytics is pricing. The KNIME Hub offers several workflows in this area[16] that researchers could use and extend, according to their needs. Geomarketing is another interesting research avenue, and the KNIME Hub offers ready implementations of the Google Maps API to help researchers automatically extract geographic coordinates from addresses and compute distances between points.[17] Such a tool might help address important questions pertaining to location marketing.

4. Summary

Empirical analyses for ML and marketing analytics have been dominated by script-based coding, on platforms such as R or Python, which offer strong performance, versatility in terms of the tasks that can be accomplished, and active community support. Yet their script-based format appeals mainly to researchers who take a written-based learning approach (i.e., reading lines and pages of code). In this article, we propose an alternative that should appeal to researchers with a preference for visual learning (i.e., a single image containing a workflow represented by configurable nodes and connecting arrows). We introduce the KNIME Analytics Platform which also offers performance, versatility and active community support, but in a visual learning environment that is designed for collaboration. We hope this article encourages ongoing efforts

[16] Choosing the right product price is critical for any company, and the many optimal pricing strategies all reflect pricing analytics. Some of them have been implemented with KNIME Analytics Platform. For an example of a workflow not including ML, see https://kni.me/w/O04qJIa1oD94M1rd; one that includes ML is available at https://kni.me/w/szC3HhTNkE12-r6h

[17] Nodes for geolocation using Google Maps are available at https://www.knime.com/book/geo-nodes

to democratize coding, as well as more interest in the constantly evolving fields of ML and marketing.

Throughout the paper we first conceptualize visual-based coding, introduce the KNIME environment (including its integration with other tools and platforms such as R, Python, Google Cloud Vision and Twitter API) and the KNIME hub for learning, sharing and reusing projects (i.e., workflows). Then, we developed an initial set of five projects that are at the intersection between marketing and ML. With these projects, we offer marketing researchers and ML enthusiasts the possibility to accelerate their learning curve in ML projects related to: Customer Churn, Consumer Sentiment, Automated Analysis of Images, SEO and Customer Experience. Furthermore, we provided two detailed step-by-step guides for the Customer Experience project, and one of the Sentiment Analysis projects. These guides can help readers to understand in a more guided fashion the steps required to build and deploy ML projects in KNIME Analytics. We hope that these projects will constitute a starting point, and that other users can share as well their workflows in the repository that we created for Marketing Analytics. This is an effort that goes into the direction of open-learning and collaboration amongst researchers via sharing visual knowledge in the form of workflows for ML and marketing.

References

Ascarza, E., Neslin S., Netzer, O., Anderson Z., Fader, P., Gupta S., Hardie., Lemmens ., Libai B., Neal D., -Provost, F., Shrift, R.. (2018), In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Customer Needs and Solutions*, 5(1), 65-81

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*, 94, 290-301.

Berthold M. (2014). Welcome to the new KNIME. KNIME Blog, 2014. Accessed on April 1, 2021, available at: https://www.knime.com/blog/welcome-to-the-new-knime

Berthold, Michael, C. Borgelt, F, Höppner, F. Klawonn, R. Silipo. (2020). Guide to Intelligent Data Science, Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

Ciechanowski, L., Jemielniak, D., & Gloor, P. A. (2020). TUTORIAL: AI research without coding: The art of fighting without fighting: Data science for qualitative researchers. *Journal of Business Research*, 117, 322-330.

Cowley, S. (2020). The YouTube SEO Project: Teaching search engine optimization through video. *Marketing Education Review*, 30(2), 125-131.

Davis, S. E. (2007). Learning styles and memory. *Institute for Learning Styles Journal*, 1(1), 46-51.

de Haan, E. and Elena M. (2020). The incremental value of unstructured data in predicting customer churn. *MSI Working Paper Series*, Report No. 20-105.

Dechant, A., Spann, M., & Becker, J. U. (2019). Positive customer churn: An application to online dating. *Journal of Service Research*, 22(1), 90-100.

Deloitte (2021). The CMO Survey 2021, accessed on March 21, 2021, available at: https://cmosurvey.org/wp-content/uploads/2021/02/The_CMO_Survey-Highlights_and_Insights_Report-February-2021.pdf

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), 82-89.

Guirao, Marc (2020), Predicting Employee Attrition with Machine Learning, KNIME Blog 2020, accessed on April 5, 2021, available at: https://www.knime.com/blog/predicting-employee-attrition-with-machine-learning

Hartmann, J., Heitmann, M., Schamp, C., & Netzer, O. (2021). The power of brand selfies in consumer-generated brand imagery. *Columbia Business School Research Paper*, Forthcoming.

Hattie, J. (2011). Which strategies best enhance teaching and learning in higher education? In D. Mashek & E. Y. Hammer (Eds.), Claremont applied social psychology series: Vol. 3. Empirical research in teaching and learning: Contributions from social psychology (p. 130–142). Wiley-Blackwell.

Heitmann, M., Christian S., Jochen H., & Christina S. 2020. More than a feeling: Benchmarks for sentiment analysis accuracy. Available at SSRN 3489963.

Herhausen, D., Ludwig, S., Grewal, D., Wulf, J., & Schoegel, M. (2019). Detecting, preventing, and mitigating online firestorms in brand communities. *Journal of Marketing*, 83(3), 1-21.

Homburg, C., Theel, M. and Hohenberg, S., 2020. Marketing excellence: nature, measurement, and investor valuations. *Journal of Marketing*, 84(4), 1-22.

Hubspot (2020), accessed on April 11, 2021, available online at: https://blog.hubspot.com/marketing/how-to-do-keyword-research-ht

KNIME (2020). Components Guides. Accessed on March 15, available at: https://docs.knime.com/2020-07/analytics_platform_components_guide/index.html

Krätzig, G. P., & Arbuthnott, K. D. (2006). Perceptual learning style and learning proficiency: A test of the hypothesis. *Journal of Educational Psychology*, 98(1), 238.

Labrecque, L. I., Patrick, V. M., & Milne, G. R. (2013). The marketers' prismatic palette: A review of color research and future directions. *Psychology & Marketing*, 30(2), 187-202.

Li, X., Shi, M., & Wang, X. S. (2019). Video mining: Measuring visual information using automatic methods. *International Journal of Research in Marketing*, 36(2), 216-231.

Li, Y., & Xie, Y. (2020). Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1), 1-19.

Liu, L., Dzyabura, D., & Mizik, N. (2020). Visual listening in: Extracting brand image portrayed on social media. *Marketing Science*, 39(4), 669-686.

Ma, L., & Sun, B. (2020). Machine learning and AI in marketing–Connecting computing power to human insights. *International Journal of Research in Marketing*, 37(3), 48, 1-504.

Marketing Tech Advisor (2019). 8 Machine Learning Marketing Certificates." accessed on March 24, 2021, available at: https://www.martechadvisor.com/articles/machine-learning-ai/8-machine-learning-marketing-certificate/

Mela, C., & Moorman, C. (2018). Why marketing analytics hasn't lived up to its promise. Accessed on March 24, 2021, available at: https://hbr.org/2018/05/why-marketing-analytics-hasnt-lived-up-to-its-promise

MIT News (2016). Brain processing of visual information. Accessed on March 25, available at: https://news.mit.edu/1996/visualprocessing

Mordor Intelligence (20201). Accessed 22-04-21, available at: https://www.mordorintelligence.com/industry-reports/ai-image-recognition-market

Nanne, A.J., Antheunis, M.L., van der Lee, C.G., Postma, E.O., Wubben, S. and van Noort, G., (2020). The use of computer vision to analyze brand-related user generated image content. *Journal of Interactive Marketing*, 50, 156-167.

Peng, L., Cui, G., Chung, Y. and Zheng, W., (2020). The faces of success: Beauty and ugliness premiums in e-commerce platforms. *Journal of Marketing*, 84(4), 67-85.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015.

Proserpio, D., Hauser, J. R., Liu, X., Amano, T., Burnap, A., Guo, T., Lee, D., Lewis, R., Misra, K., Schwarz, E., Timoshenko, A., Xu, L., & Yoganarasimhan, H. (2020). Soul and machine (learning). *Marketing Letters*, 31(4), 393-404.

Rocklage, M. D., Rucker, D., & Nordgren, L. F. (2018). The Evaluative Lexicon 2.0: The measurement of emotionality, extremity, and valence in language. *Behaviour Research Methods* 50 (4), 1327-1344.

Schweidel, D., Martin R., & Thomas R. "Supporting Content Marketing with Natural Language Generation," Marketing Science Institute Working Paper Series 2021.

Thontirawong, P., & Chinchanachokchai S. (2021). Teaching artificial intelligence and machine learning in marketing. *Marketing Education Review*, Forthcoming.

Timoshenko, A., & Hauser, J. R., (2019). Identifying customer needs from user-generated content. *Marketing Science*, 38 (1), 1-20.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, 51(4), 463-479.

Tursi, V., & R. Silipo (2018). From Words to Wisdom: An Introduction to Text Mining with Knime.

Verboven, S., Berrevoets J., Wuytens J., Baesens, B., & Verbeke, W. (2020), Autoencoders for strategic decision support. *Decision Support Systems*, Forthcoming.

Villarroel Ordenes, F., and Zhang S. (2019). From Words to Pixels: Text and Image Mining Methods for Service Research. *Journal of Service Management*, 30 (5), 593-620.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. Proc. of HLT-EMNLP-2005.

Yen, J. C., Lee, C. Y., & Chen, I. J. (2012). The effects of image-based concept mapping on the learning outcomes and cognitive processes of mobile learners. *British Journal of Educational Technology*, 43(2), 307-320.

Zhang, S., Friedman, E., Zhang, X., Srinivasan, K., & Dhar, R. (2020). Serving with a Smile on Airbnb: Analyzing the Economic Returns and Behavioral Underpinnings of the Host's Smile, Available at SSRN (2020).

Zhu, B., Baesens, B., & vanden Broucke B., (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information Sciences*, 408, 84-99.

Table 1: Comparison of Analytic Platforms in Relevant Areas Related to this Manuscript

Article	KNIME	RAPIDMINER	SAS	PYTHON	R	ALTERYX	AZURE ML
Open Source	✓	✓*		✓	✓		
Visual Programing	✓	✓	✓			✓	✓
Community Repository	✓			✓	✓		
Open Architecture	✓			✓	✓	✓	

*Different alternatives depending on type of Rapid Miner (Studio or Go), and user characteristics (academic or not)

Table 2: Curation of ML for marketing in the Knime Hub. Some case studies contain two workflows: one for training and one for deployment.

Case Study	Description	Additions in KNIME	Dataset	Relevant Marketing Literature	Link to Examples
Customer churn	Supervised ML. Traditional ML approach using random forest	None	Kaggle TELCOM churn	de Haan and Menichelli (2018); Amin et al. (2020)	https://kni.me/w/9cBJzpQEeZyMtSNa https://kni.me/w/PRrwnq5kb6UYXJuP
Consumer sentiment analysis	Supervised ML. Traditional ML including using support vector machine (SVM)	Twitter API*, Text processing extension	Kaggle Airline Tweets and live scraping of tweets	Heitmann et al. (2020); Kübler et al. (2020); Villarroel Ordenes et al. (2020)	https://kni.me/w/KvOyB7vz9Dt1MZOQ https://kni.me/w/Qvn6ezkH4z9KaLn-
Consumer sentiment analysis	Deep neural networks, includes social media scraping with Twitter API	Twitter API, KNIME Deep Learning Keras Integration	Kaggle Airline Tweets	Heitmann et al. (2020)	https://kni.me/w/A5JrR6DA9zxLZD9m https://kni.me/w/C80at_YXQ6PK-R19
Consumer sentiment analysis	Deep neural networks, using transformer methods with Google BERT	Twitter API and Google BERT	Kaggle Airline Tweets	Heitmann et al. (2020); Alaparthi and Mishra (2021)	https://kni.me/w/3VGz0gdnmIy-DfT4 https://kni.me/w/myYBJS_VRYmt8h2P
Consumer sentiment analysis	Lexicon-based (dictionary) sentiment analysis	Text Processing Extension	Kaggle Airline Tweets	Herhaussen et al. (2019); Heitmann et al. (2020).	https://kni.me/w/zHAUMcOElRy20qO1 https://kni.me/w/djkzPLAKhP-j3Y17
Social media brand content (image analysis)	Implementation using Google Cloud Vision API	Google Cloud Vision** and Python Integration	Brand Instagram posts	Colors and labels: Li and Xie (2020), smiles: Zhang et al. (2020)	https://kni.me/w/FEczzB1FQBnPrCQg
General workflow for image classification	Image classification using deep learning CNNs	KNIME Deep Learning Keras Integration Image Proc. Extension	Kaggle dogs vs. cats	Selfies: Hartmann et al. (2020), brand attributes: Liu et al., (2020)	https://kni.me/w/8XFC5HmWmuLd6LW1 https://kni.me/w/P8j6x0NqLsSZWkSa
CX and Topic Models	Implementation using LDA	R Integration Text Mining Extension	Customer reviews	Tirunillai and Tellis (2014); Netzer et al. (2019)	https://kni.me/w/zXa_WBQgRZz4nqq6
Search engine optimization (SEO)	Extracts common [co-occurring] words, keywords, and topics.	Google Analytics and Twitter API	Tweets and SERPs	Cowley, (2019), Hubspot, (2020)	https://kni.me/w/tST9mpGcmDVL3y_P

*Users need a Twitter API (with credentials) through the Twitter developer's page https://developer.twitter.com/en/apply-for-access
**Users need a google vision API and download its JSON file with the service account key https://cloud.google.com/vision/docs/setup

Figure 1: KNIME workbench within KNIME Analytics Platform. Image provided by KNIME.

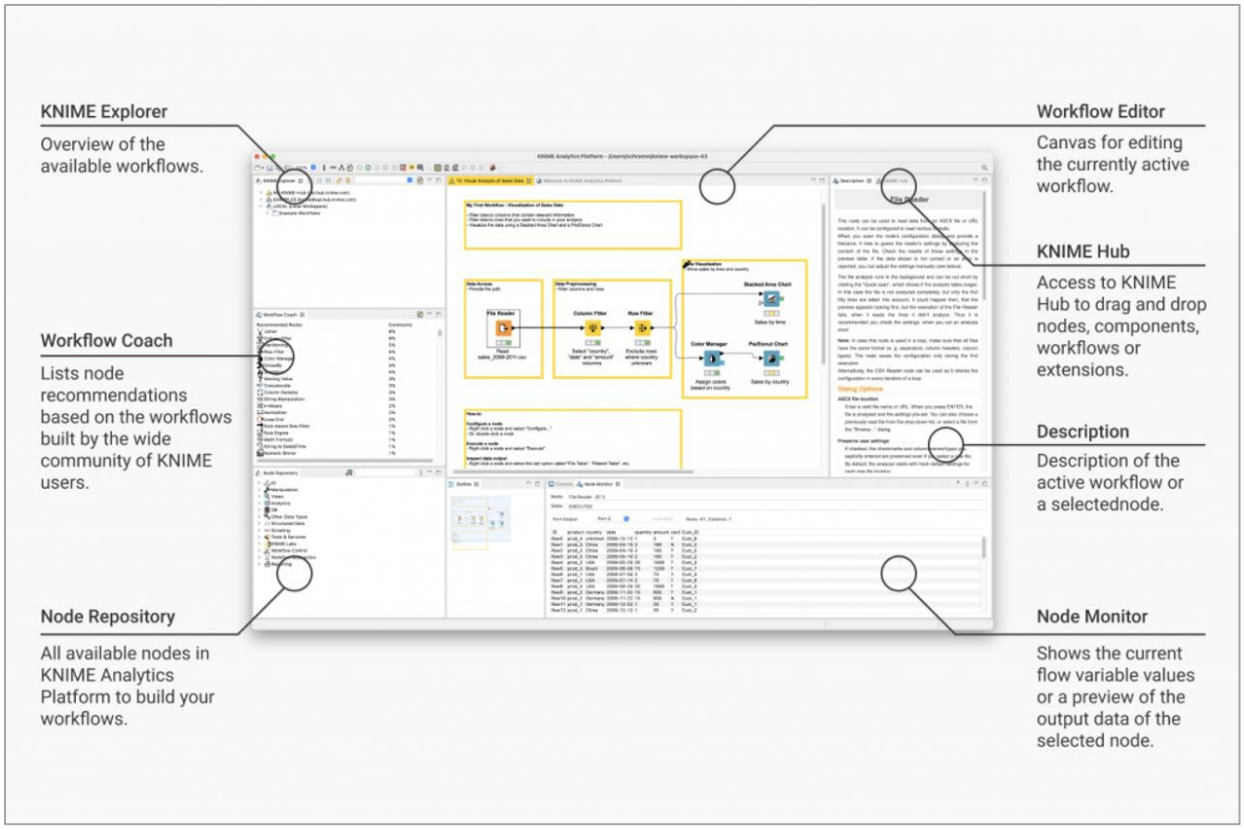


Figure 2: Node status displayed as a traffic light under each node. Image provided by KNIME.

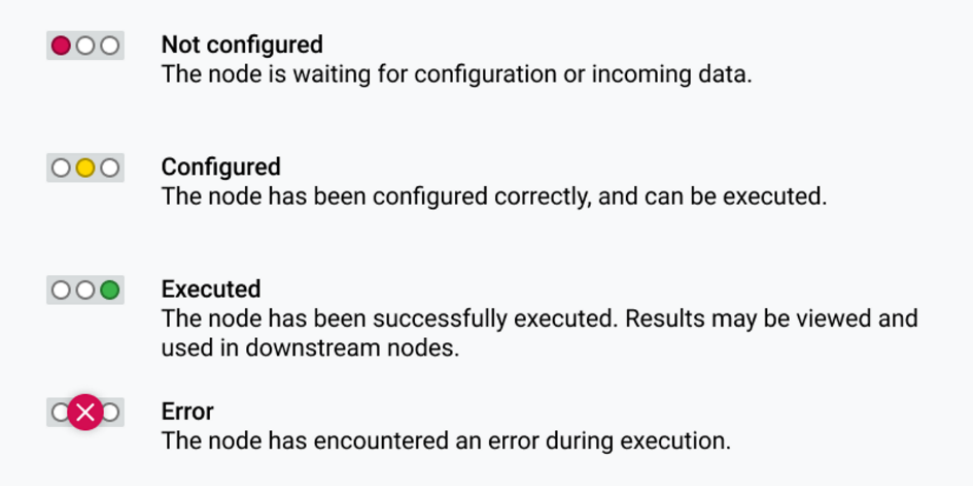


Figure 3: Nodes and components

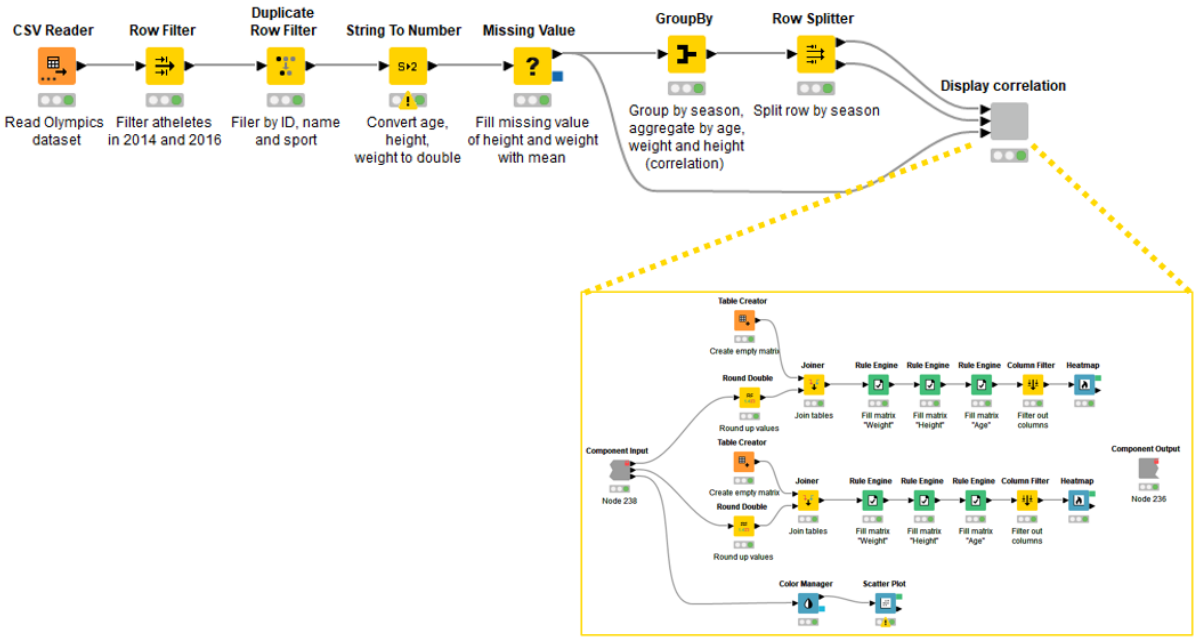


Figure 4: Common data science techniques used in digital marketing and available in KNIME Analytics Platform

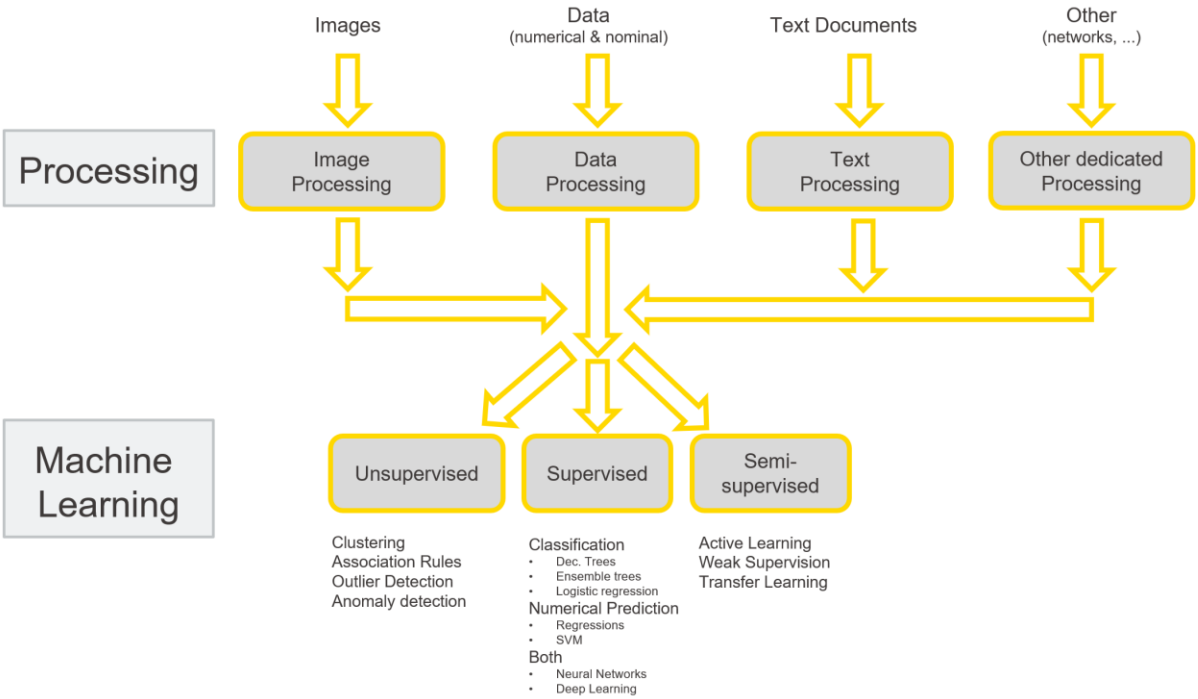


Figure 5. Training workflow to train and test a random forest algorithm for churn prediction

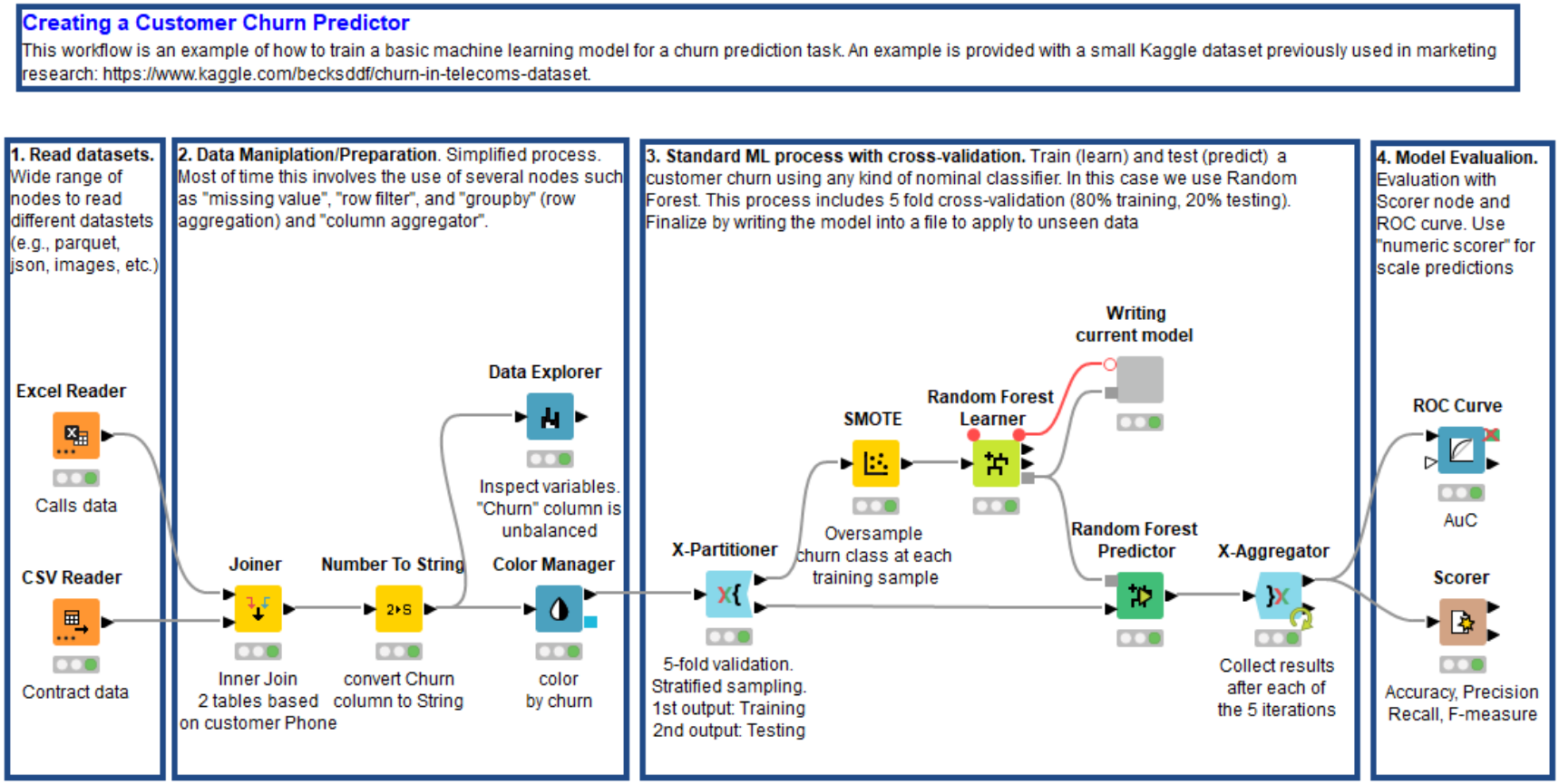
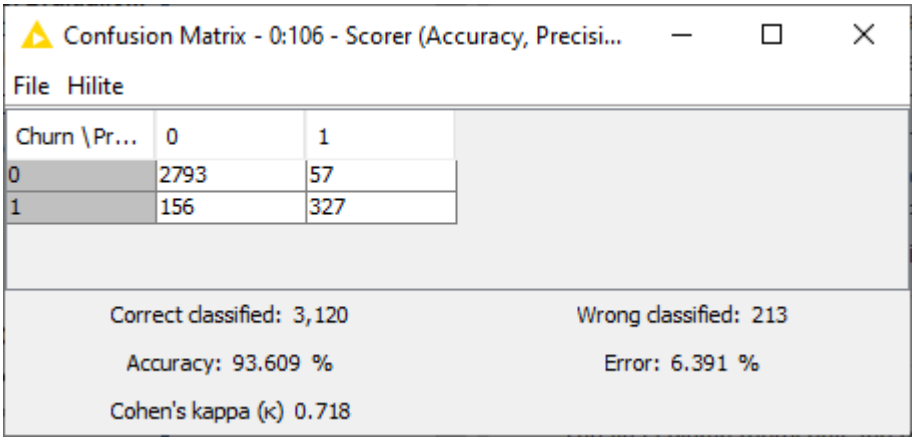


Figure 6. Scorer node

a. Summary of evaluation results



b. ROC curve

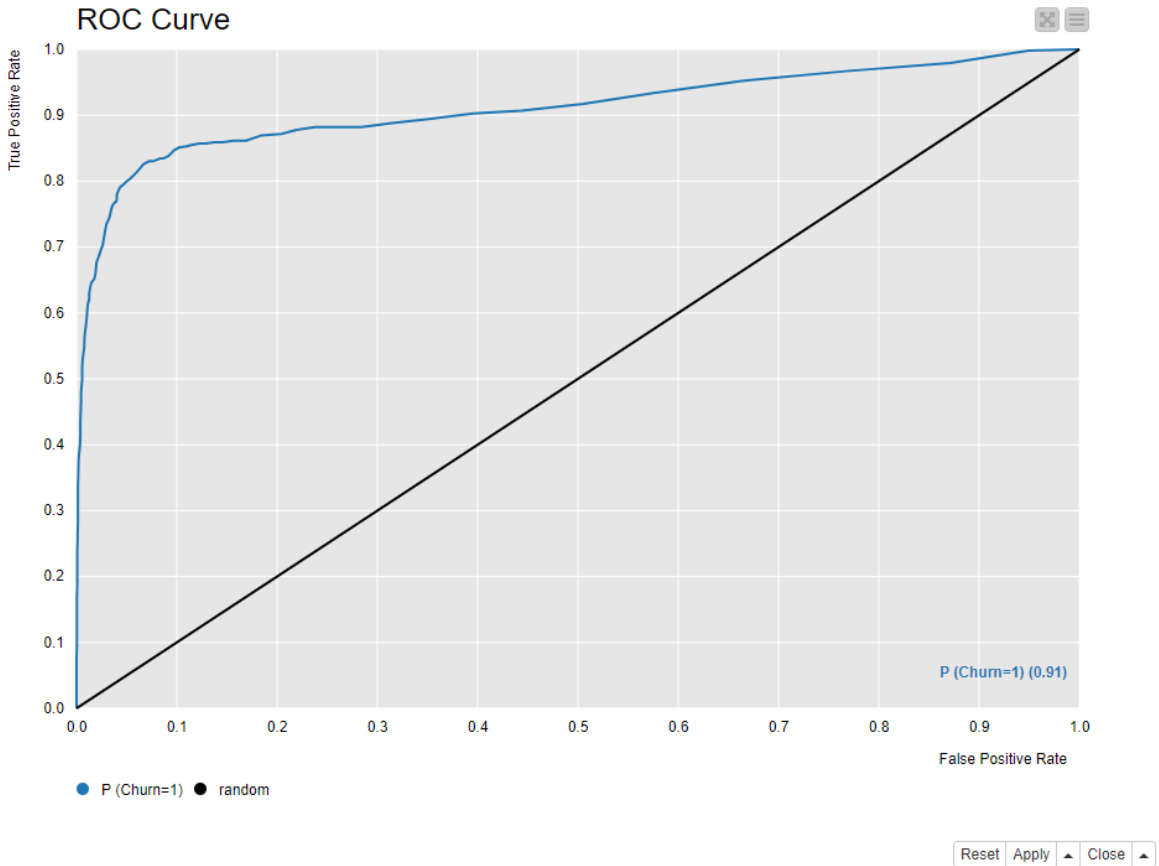


Figure 7a. Workflow for model deployment and visualization

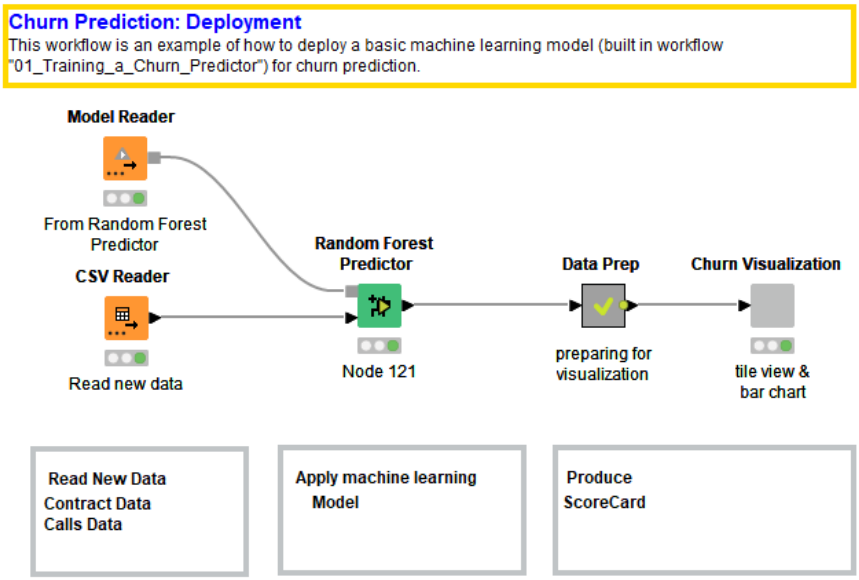


Figure 7b. Composite view of the churn visualization component (interactive)

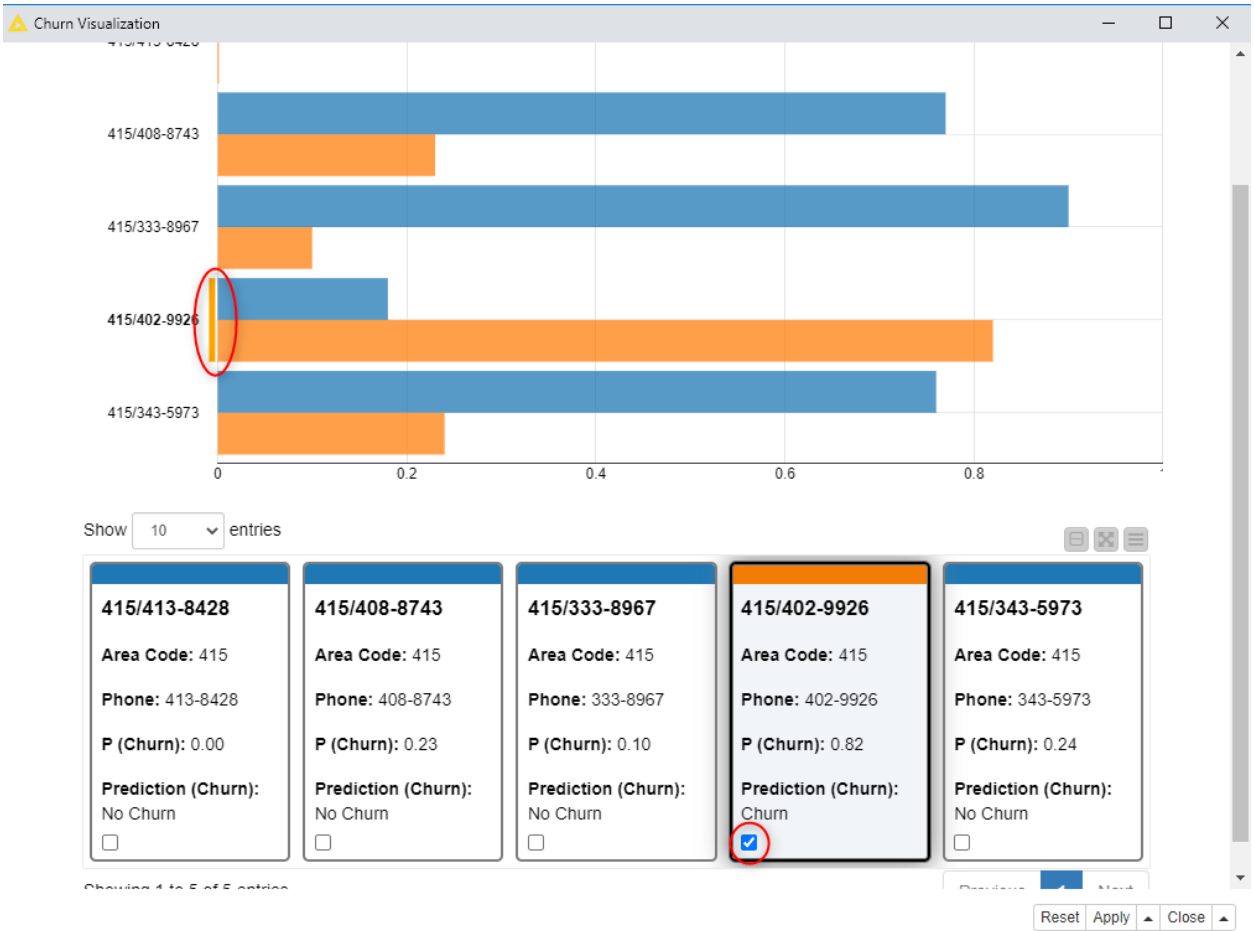


Figure 8a: Sentiment workflow

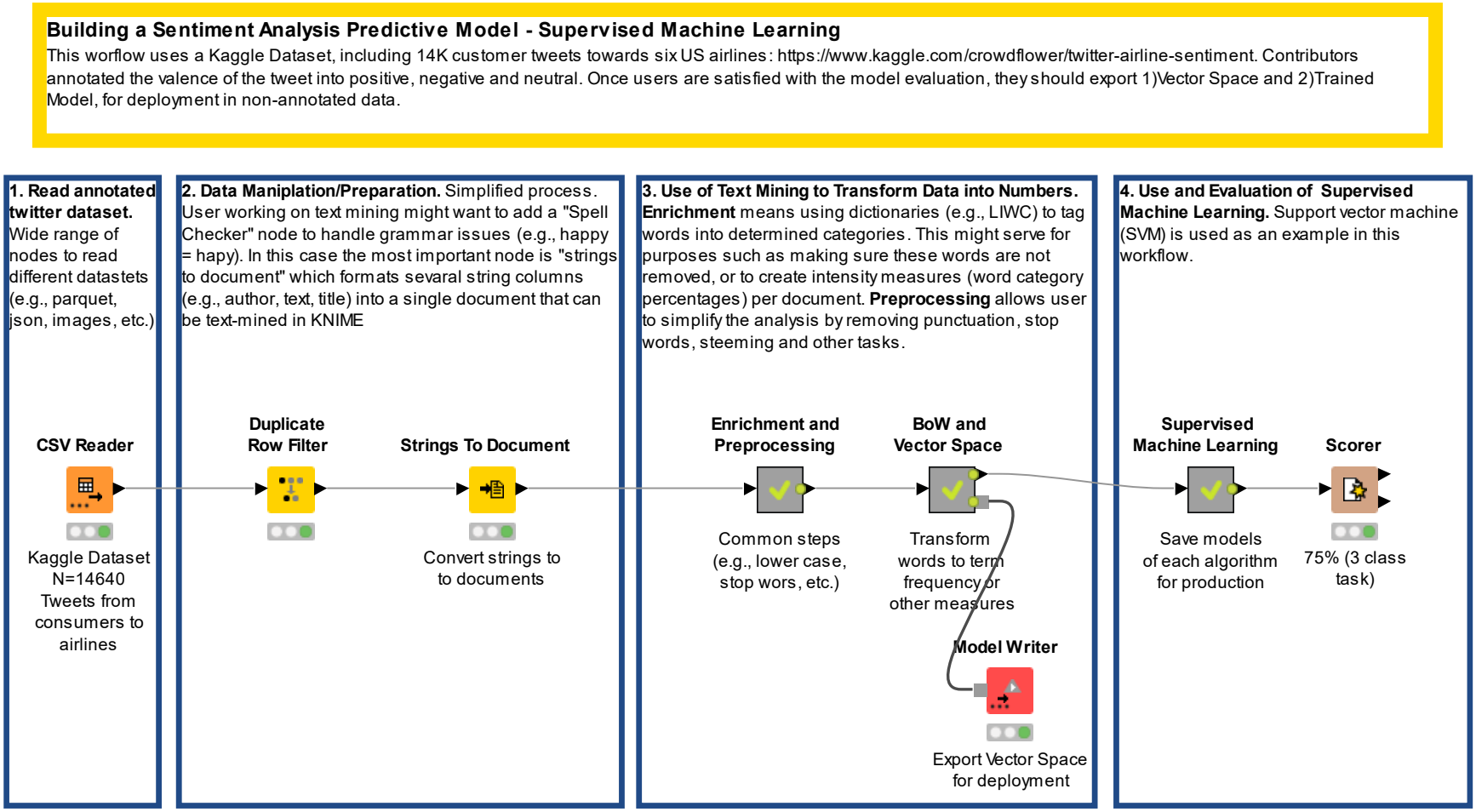


Figure 8b: Metanode enrichment and preprocessing

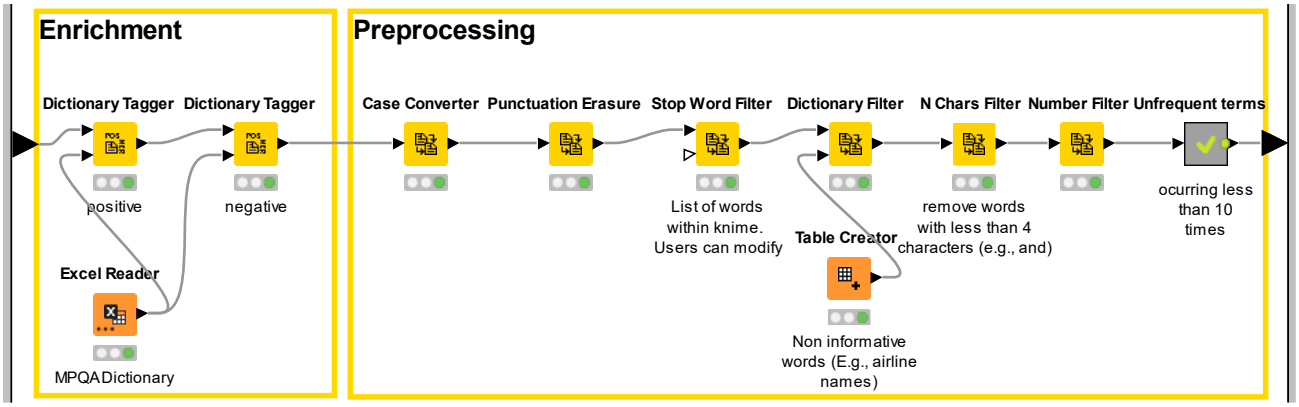


Figure 8c: Metanode bag of words and document vector

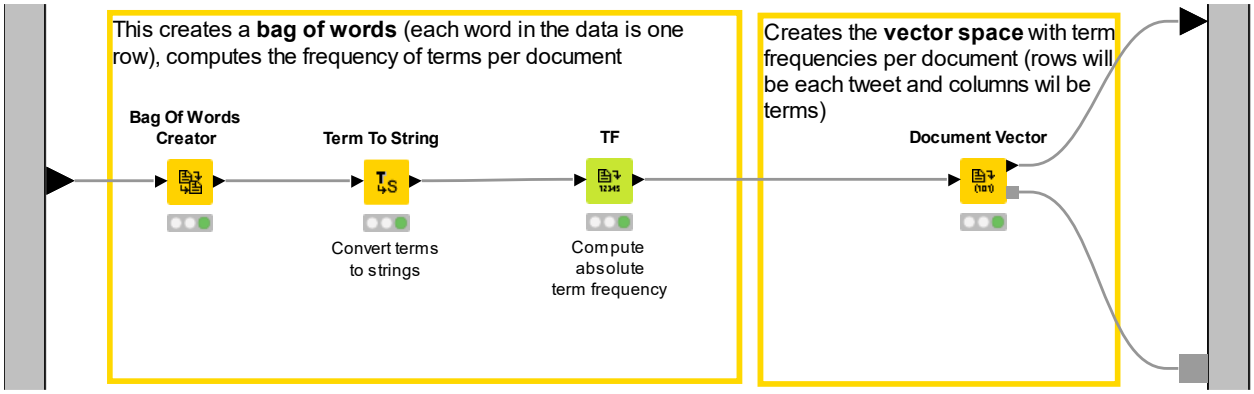


Figure 8d: Metanode supervised machine learning

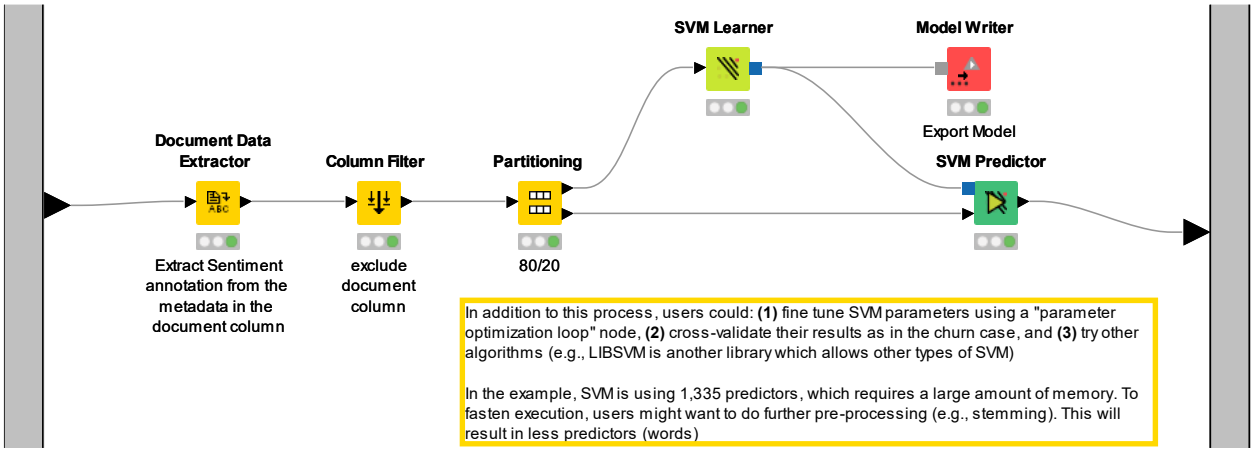


Figure 9: Deployment of sentiment predictor

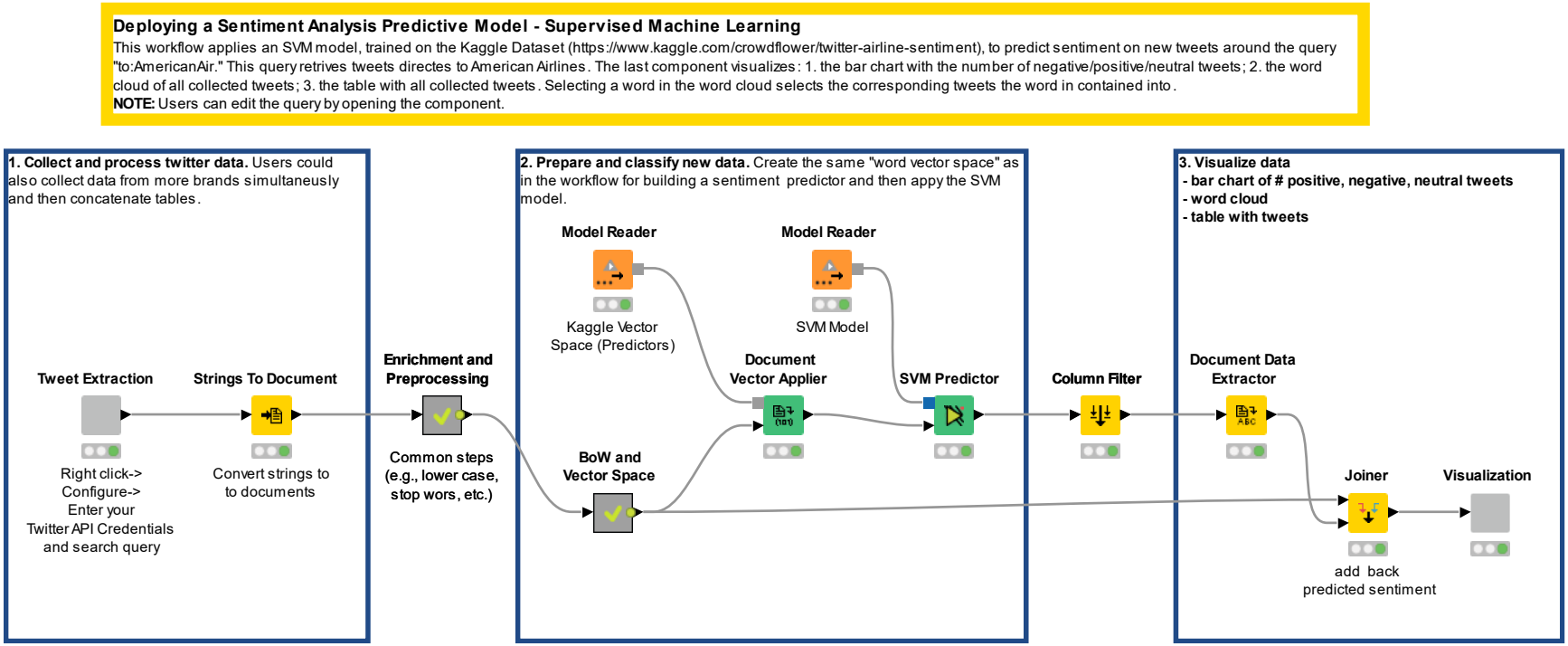


Figure 10: Visualization of tweets with estimated sentiment (red = negative, green = positive, light orange = neutral)

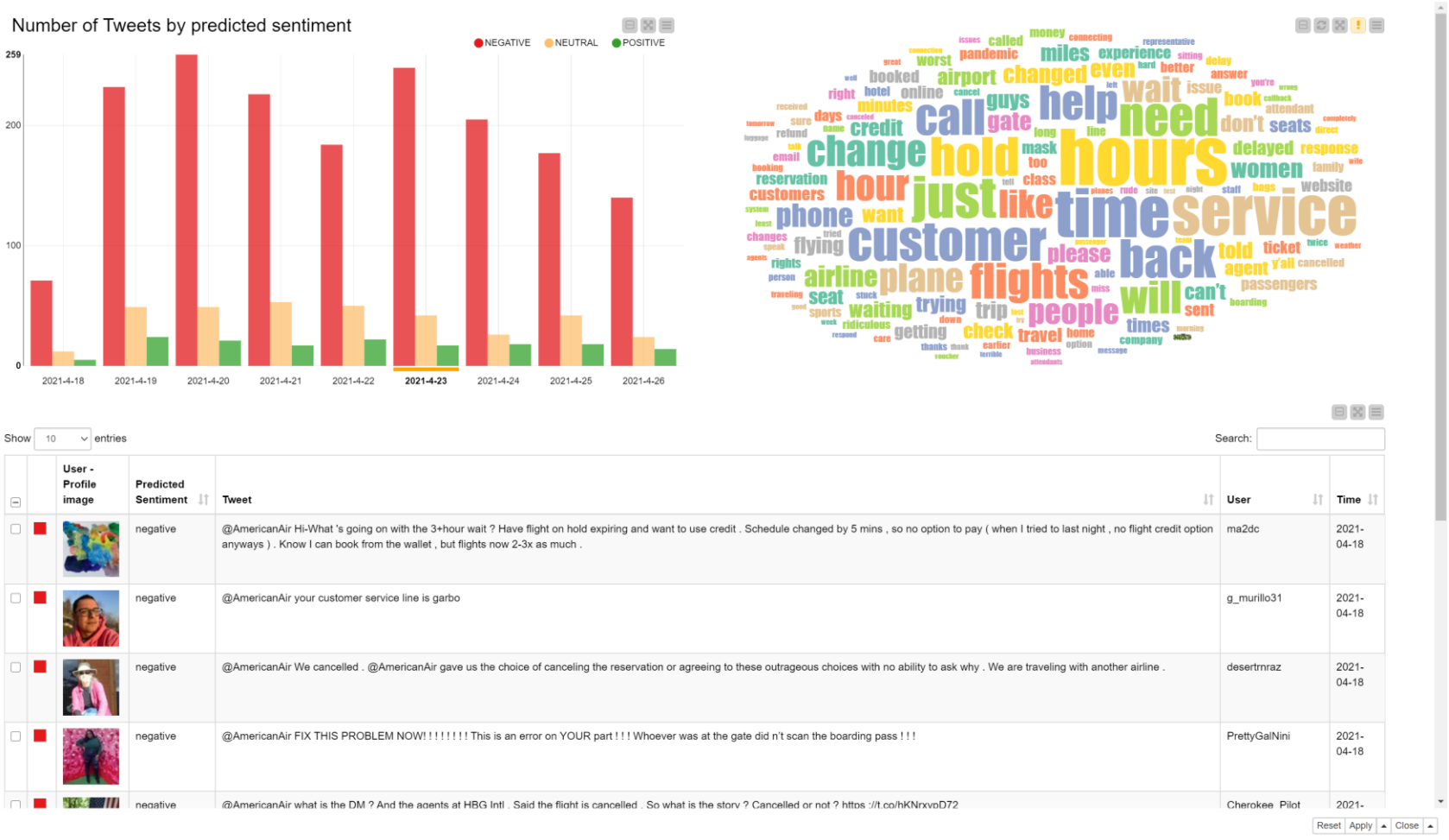


Figure 11a: Workflow to discover topics in feedback texts

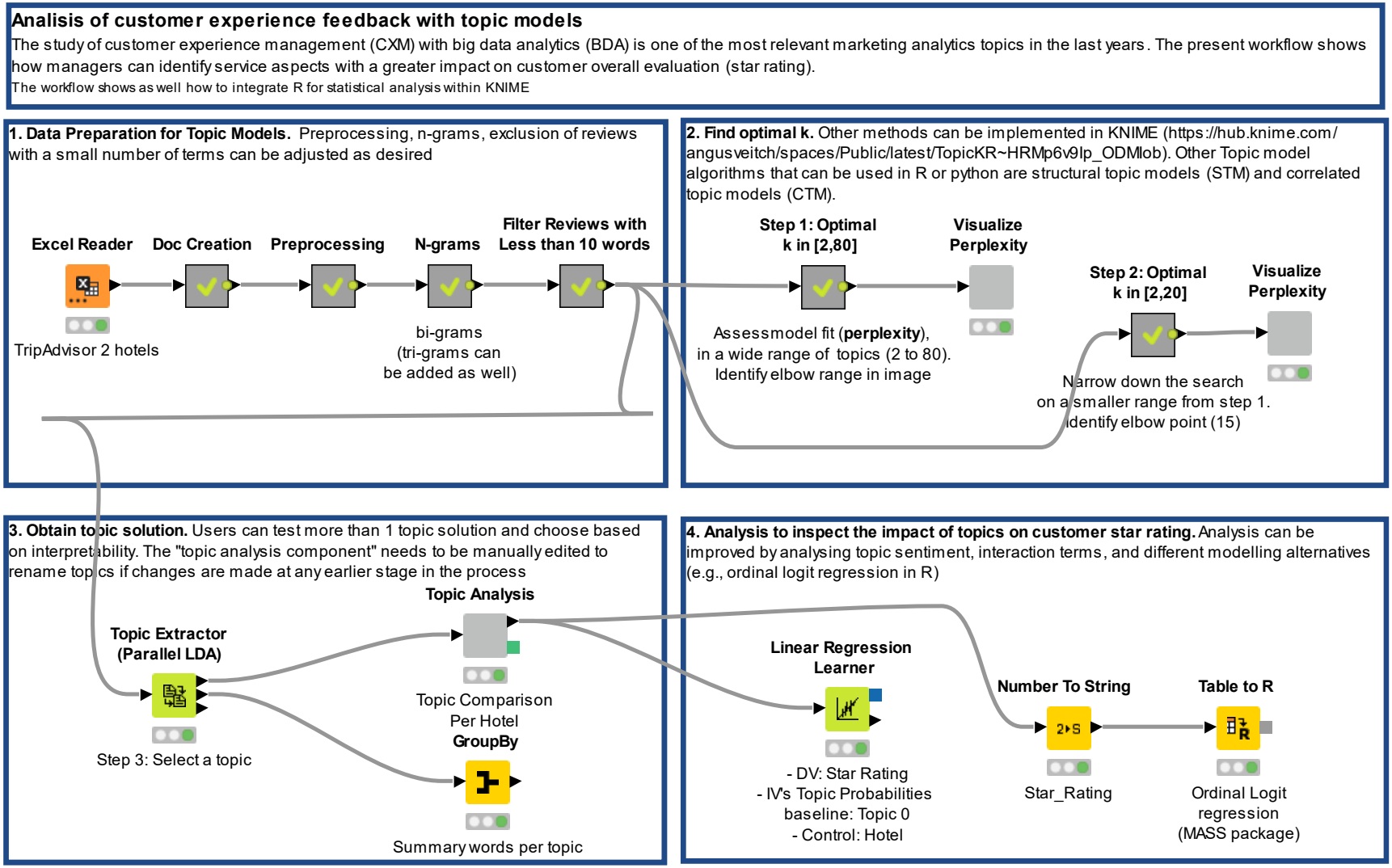
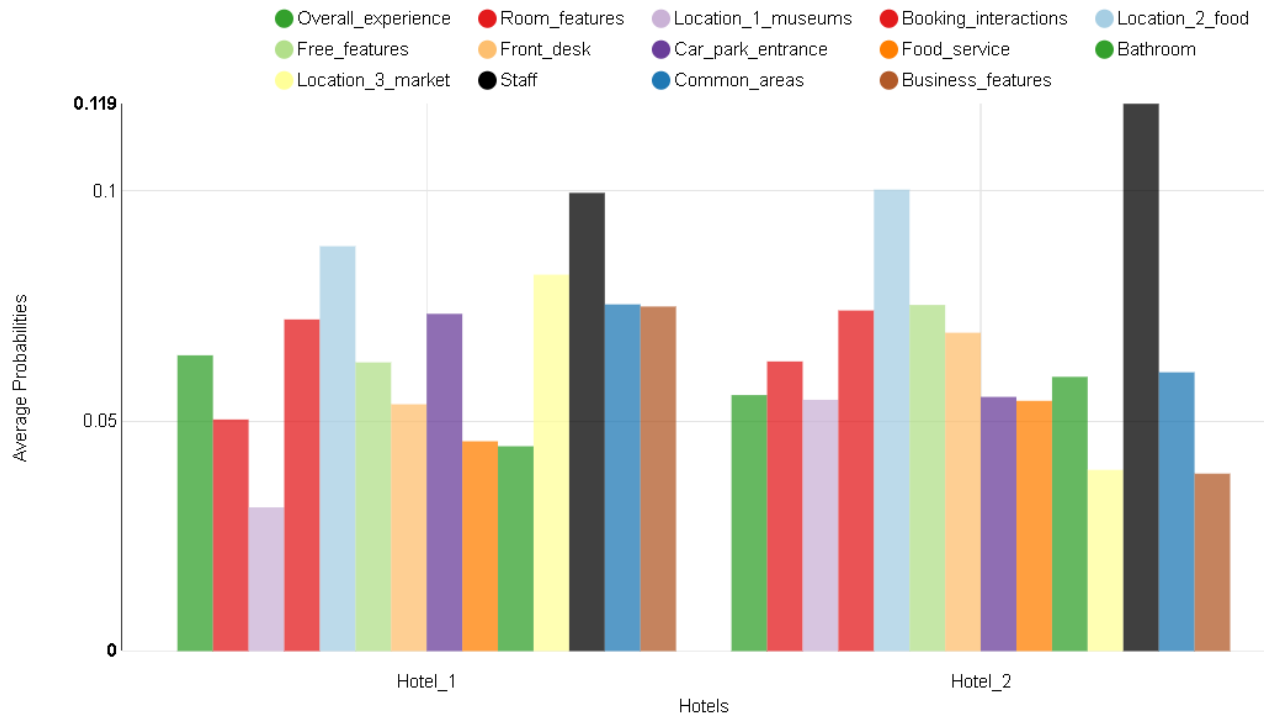


Figure 11b.Discovered topics in Customer Reviews and their Impact on the Star Rating



Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
Free_features	-2.8582	0.2656	-10.7631	0.0
Location_1_museums	-0.8639	0.2772	-3.1166	0.0018
Staff	1.1995	0.2549	4.7058	2.60E-6
Bathroom	-5.8854	0.245	-24.0239	0.0
Overall_experience	-5.5674	0.278	-20.0291	0.0
Room_features	-3.4776	0.2658	-13.0853	0.0
Booking_interactions	-5.6904	0.2247	-25.3259	0.0
Location_2_food	-0.3026	0.2599	-1.1643	0.2444
Front_desk	0.3753	0.2674	1.4034	0.1606
Car_park_entrance	-3.6104	0.2581	-13.9911	0.0
Food_service	-3.0546	0.2684	-11.3821	0.0
Location_3_market	-0.5597	0.2627	-2.1305	0.0332
Common_areas	-1.3716	0.277	-4.9521	7.60E-7
Business_features	-2.0885	0.2764	-7.5554	5.00E-14
Hotel_Name=Hotel_2	0.0089	0.0285	0.3106	0.7561
Intercept	5.7576	0.1787	32.2131	0.0

Multiple R-Squared: 0.4896
Adjusted R-Squared: 0.488

Figure 12a: Workflow to extract color dominance & object recognition from Google Cloud API

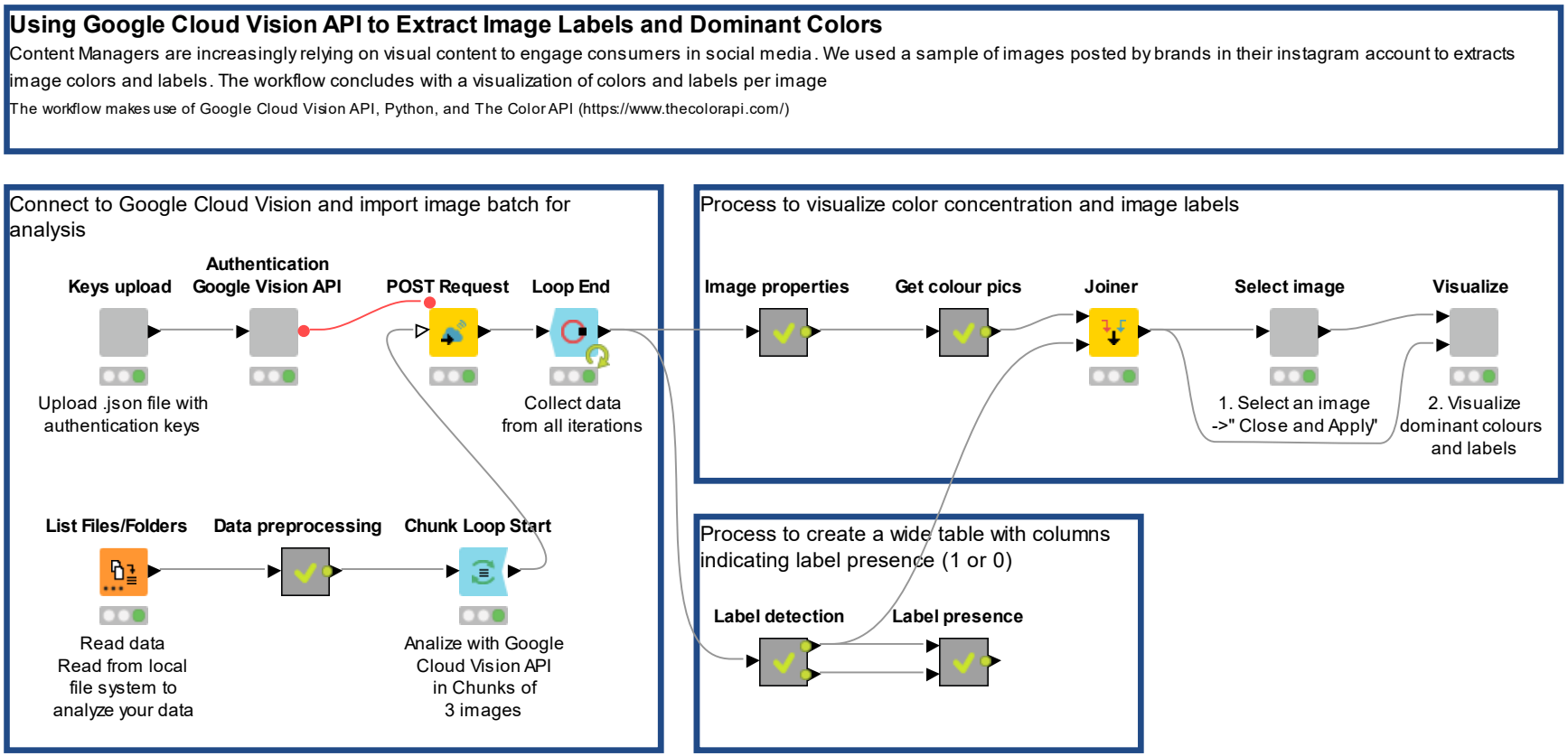
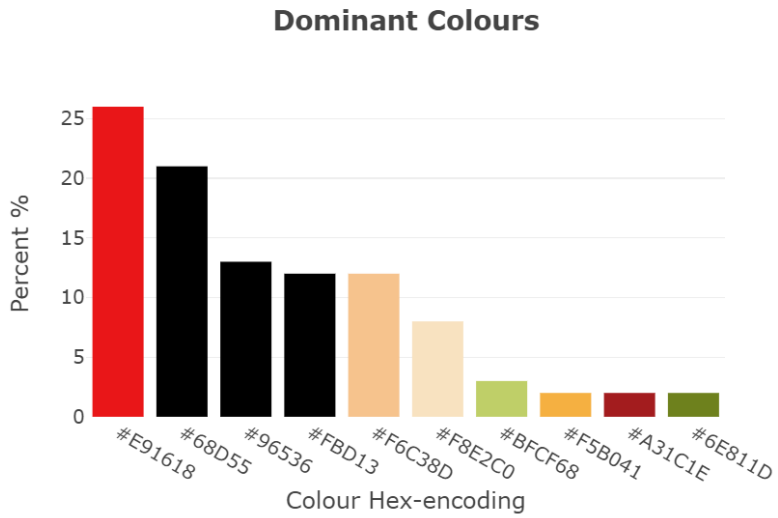


Figure 12b: Summarizing color dominance & object recognition from input image

Input image:



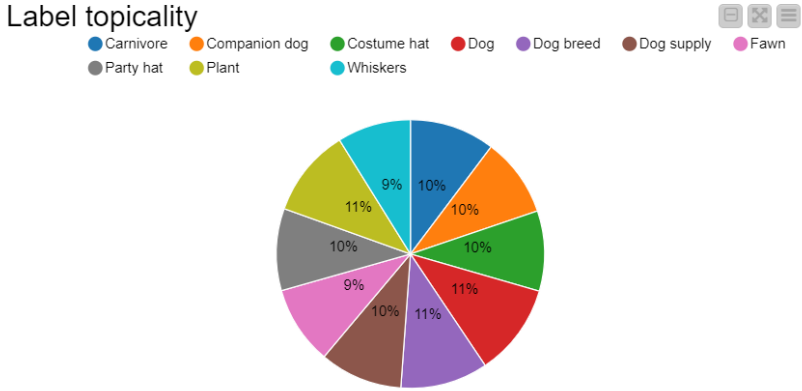


Figure 13a. Workflow for semantic keyword research for SEO

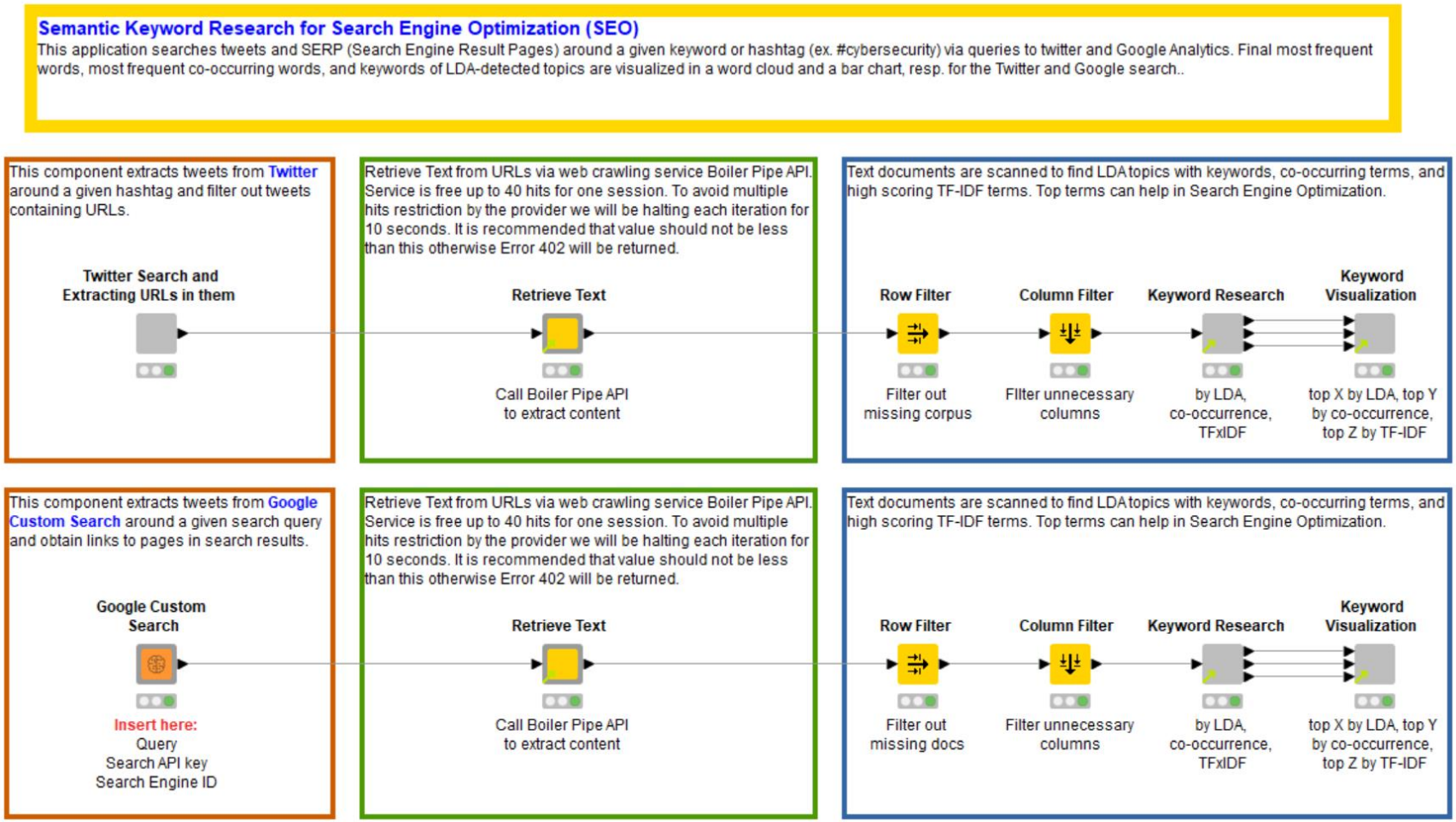
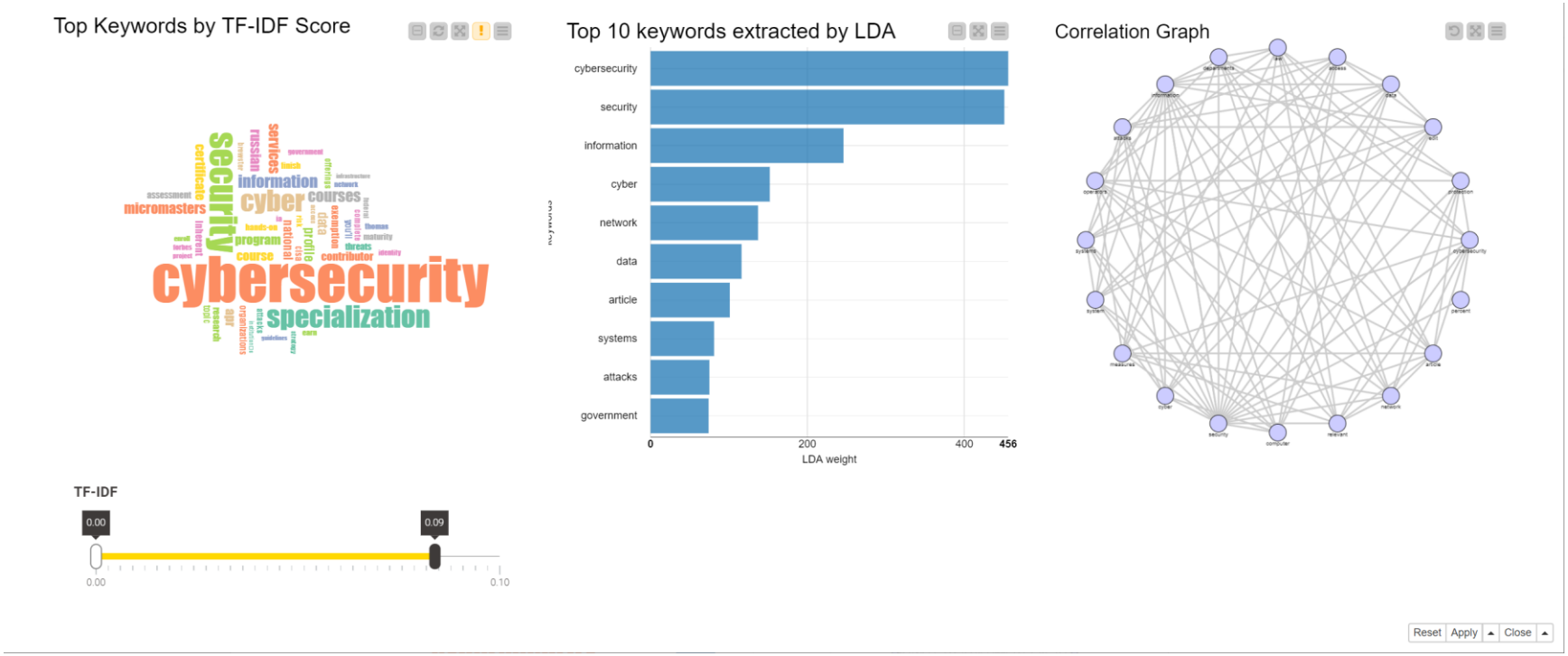



Figure 13b. Word cloud with top TF-IDF words, graph with top co-occurring words, and topic-related keywords from Google SERP around search term “cybersecurity”