Nicholas Catani

ISM 6359 – Data Mining Exam/Reflection (150 pts)

DIRECTIONS: Answer any **FOUR** (4) of the six possible questions below.

1. Reflect on the CRISP-DM Framework/Process; explain each of the phases to somebody who has no understanding of what predictive analytics can do for them. (25%)

CRISP-DM stands for cross-industry standard process for data mining. In essence, the most common framework adopted throughout the industry, and it is employed as a guideline for DM projects. It has six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

BUSINESS UNDERSTANDING: This phase helps the scientist clinch the objectives and customer's needs tied to a project. Cost-benefit analysis, selection of technologies or tools, and risk and resource evaluations are all assessed here. A strong foundation plays a vital role in the success of a DM project.

DATA UNDERSTANDING: It drives to identify, gather, and inspect the dataset that an insider's looking for to achieve the outcome. It is important because supports mapping questions such as, does it suit the company's needs? Primary or secondary data? Is it trivial to retrieve or does it require considerable financial effort and time to collect that piece of information? Does the dataset show quality issues?

DATA PREPARATION: Generally, 80-90% of the time working on a project is spent here. It is where the dataset gets massaged before getting ready for modeling. At this stage, the insider can clean the data from erroneous values, replace missing ones, determine attributes to select as labels before feeding the ML model, and apply feature engineering "a process consisting in the creation of new attributes by combining existing ones", and reformat data "an example would be nominal to numeric".

MODELING: Here a scientist selects, builds, and assesses machine learning algorithms "decision tree, random forest, naïve bayes, k-nn, linear regression, deep learning, etc." to find the best predictive model according to the dataset. Generally, depending on the approach, data might need to be split into training and test data.

EVALUATION: it has a more business perspective rather than a technical one. This phase gives insights into which model meets the company standards to achieve its strategic goals and what to do next when choosing the right one. Here, decision-makers can proceed to deployment "if satisfied", continue iterating "apply models till finding a good one", or start a new project.

DEPLOYMENT: It starts with generating a report to document what went well, what could have been better, and how to improve in future projects - the team might create a presentation of data results if those must be pitched to stakeholders. As regards the chosen model, it will be deployed for prediction purposes in the company.

2. Much of your time building predictive analytical models consists of data preparation. List and explain many of the common data preparation functions that you might perform on your data sets to get the ready for the machine learning algorithms. Be as specific as possible. (25%)

Data preparation is the process where raw data gets manipulated and brought to a decent form for further analysis – data is commonly created with missing values, irrelevant attributes, and other errors. Therefore, it represents a good practice to avoid killing a sensitive algorithm. Tasks such as collecting, cleaning, or transforming data fall into such a phase. One of the benefits of preparing data I may recall is, perhaps enduring the data analysis to produce reliable results, so it can support business executives in making better strategic decisions. Generally, the process consists of multiple steps before moving forward.

DATA COLLECTION: data is collected from internal or external systems, or data warehouses, by constantly keeping in mind whether is relevant to the project goals.

DATA CLEANSING: once the dataset has been explored and faulty data identified, the scientist proceeds by removing or fixing errors, correcting typos or misspellings, filling in missing values, removing irrelevant attributes, and harmonizing inconsistent entries. In RapidMiner, Professor LaBrie has shown us various operators that fit the purpose. For example, "Select attribute" allows a user to select relevant columns for further analysis. "Replace missing values", is useful for missing data. Some algorithms are extremely sensitive to missing values and are not so forgiving, deep learning for instance. Another operator I might think of is "detect outliers". In statistics, an outlier describes a single observation that is located abnormally away from clustered values in a random sample of the population. In Data Science is important to remove outliers because they might be measurement errors or poor sampling. “Detect outliers” can remove those pieces of data that might confuse the algorithm, thus compromising the prediction.

FEATURE ENGINEERING: it may involve the creation of new fields or attributes aggregating values from existing attributes. In RapidMiner, the operator responsible for this task is named “Generate Attribute”.

DATA SPLITTING: refined data is split into more subsets. One part is used for evaluating or testing the data “testing data” and the other for training the model “training data”. Typically, this process is done to avoid a problem called overfitting. It is an instance where a selected model suits the train data of a dataset perfectly but fails to fit other datasets. Also, it is worth paying attention to the difference between cross-validation and split data – operations found in RapidMiner as well. As mentioned above, both are trade-offs between the amount of data used for training the model and the amount of data used for testing, However, in cross-validation, the split is done into k-folds where the train occurs in all partitions except one, left for evaluation. It is iterated for k times, with a different partition reserved for data testing each time. Once done, the technique extracts the accuracy average from the sum of all folds. Whereas in split data, there’s no iteration. Training and testing data is partitioned and run just once. Of the two, cross-validation is the most reliable because of the ability to detect overfitting and reduce potential bias that might confuse the ML model.

DATA COMPRESSION: in certain datasets, the scientist deals with attributes that contain a wide range of numbers or values. Therefore, finding a way to compress them, would benefit the model in the long run. Besides, many algorithms perform better when input variables have a standard probability distribution – to obviate, it could be helpful to “Discretize” or “Normalize” variables. The former transforms continuous values sitting in an attribute into discrete ones. Whereas the latter transforms everything into a percentage. Notice, "normalize" is very appropriate for those models which required all variables to be numeric such as regression models "logistic & linear", otherwise it would choke.

DATA LABELING: no supervised model can be executed without labels. They must be informative, perceptive, and independent if a scientist wants to farm a good model. RapidMiner has a feature called "set role" where a user can select the interested attribute to select has a label to check its predictions for accuracy.

3. Explain/define the following two data mining models: classification models and regression models. Explain how they differ. Where would you use one over the other? Please give examples. (25%)

Classification models are a set of supervised machine learning algorithms. They are designed to read data input (label) and transform it into an output while classifying inputs. One of the very first examples of a classification model I may recall is; a user email has an option where incoming emails can be classified either as spam or not, for example.

Decision tree, Random Forest, K-nearest neighborhood, naïve Bayes, deep learning, support machine vector, or artificial neural network "remind last three require more effort in data preparation for a scientist because unforgiving toward data errors or missing values" are all great instances falling into this category – supervised & classification models. Whereas regression models, even though supervised ones as well, they try to understand the correlation between different independent variables and a targeted outcome. Examples of models are Linear regression and Logistic regression. A simple way to explain it is by using Cartesian graphs. The scientist selects an independent variable – on the X-axe – and checks, whether that variable will most likely or unlikely, will have an impact on the dependent one – represented on the Y-axe – by using the Pearson linear correlation index.

All that aside, there are two major differences to keep in mind: the first distinction is between ***the classification model and the regression model.*** The outcome extracted from the predicted label made with a classification model is generally in binomial form – whether is a yes or no type of answer. Straight example, a scientist is trying to predict either a new apartment complex recently built in Capitol Hill, will be sold or not. In the regression model, in contrast, the prediction tends to boil down to a specific numeric value. Linear regression, for example, can be extremely powerful to predict the future stock price of a public company or the selling price of a house. The second distinction is between ***linear regression and logistic regression.***

While the former has been already mapped above where the models help to narrow down the analysis toward a single specific number. The logistic regression model, rather than focusing on predicting a single value, falls back on the classification group "mentioned at the beginning" and tries to understand why a singular event occurred.

4. Describe what clustering is used for. What are some of the different methods/algorithms for clustering analysis? (25%)

Clustering is a technique applied to unsupervised machine learning algorithms where the scientist does not have label data to make a prediction – no answer at his disposal. Generally, it works as a way of grouping data points into different clusters of data points all alike. It does accomplish that by finding patterns in the unlabeled data such as color, or size, and splits them accordingly. As Professor LaBrie mentioned in class, clustering analysis can be performed in Marketing analytics to partition targeted consumers into different categories depending on their age, sex, purchasing behaviors, interests, wealth, and so on. – marketing segmentation.

It is interesting to notice that in unsupervised learning, there's an additional technique called dimensionality reduction. Conversely to clustering, this technique helps the scientist to deal with a massive quantity of observations and sparse data, by compressing high-dimensional spaces into low ones, which otherwise would make data analysis unfeasible.

With that aside, Clustering is an unprecise technique. But, good enough if the scientist uses it as a pre-processor for grouping data to employ afterward, as labels in supervised learning. As far as I'm concerned, there are four types of approaches: Prototype-based clustering, Density clustering, Hierarchical clustering, and Model-based clustering.

- *PROTOTYPE-BASED CLUSTERING:*

In this approach, an observation tends to be assigned to its nearest prototype – typically called centroid if the scientist is dealing with numeric data, otherwise medoid if categorical. The simple examples that fall into this category are, **K-MEANS** & **K-MEDOIDS**. There are some important differences to highlight between the two; **k-means** clustering is efficient if employed for small datasets. It does not support categorical values "like regression models" and a random point can fall into white space. Conversely k-medoid, the model needs a mass amount of data to position the centroid, without having the possibility to land on a random white space to start clustering – it must be a data point.

- *DENSITY-BASED CLUSTERING:*

Here the intuition is based on a cluster in a generic data space where it has a juxtapose area of high density, detached from sparse clusters regions. This is a good approach to employ when the dataset contains irregularities, noises, and outliers. An example is **DBSCAN**.

- *HIERARCHICAL CLUSTERING:*

The creation of clusters comes by having a predominant order from top to bottom. An example of this kind might be the application software we have installed on our computers. Files are stored in folders, folders in another folder, all the way up to local disk > this pc. A hierarchical system.
Subsequently, this technique is divided into agglomerative and divisive clustering.

- *MODEL-BASED CLUSTERING:*

As mentioned in its name, this form of clustering start with the assumption of the data comes from a predefined model and tries recovering that model from the data.