Anav Prasad
CSCI-GA 3033 061: Professor Anasse Bari
June 15, 2022

Predictive Analytics Quiz 1

Question 1. Define in your own words: CRISP-DM (12pts) and explain all its phases.

Solution:

CRISP-DM, which stands for CRoss Industry Standard Process for Data Mining, is a method for the course a predictive analytics project, or any Data Analytics Project, might take in its lifetime. The phases for the same are as follows:

- Business Understanding: The first phase focuses on, first of all, understanding the problem statement as to how it corresponds in the business domain and to properly understand the objectives of the project. Furthermore, it prioritizes the goals that one might expect out of the project and a roadmap / project plan to achieve the same.
- Data Understanding: The second phase relates to the data collection, analysis and comprehension. That is to say, the analysts should acquire their data, analyze its format and features, understand it by using various techniques (be it visualized or otherwise) and, finally, verify its correctness and quality.
- Data Preparation: The third phase is about preparing the data that is going to be used in the training of the model. This includes selection of data (that is selecting the datasets to be used and documenting why or why not), cleaning of the data (removal of noise, excess dimensions, normalization where needed etc), construction of data (formation of proper attributes with introduction of new ones as per need), integration of data (combination of related datasets for simplification and integration), formatting of data (final round of changes to the data, like string to int, decimal precision of float values etc, so as to perfect it for the modeling phase).
- Modeling: The fourth phase refers to the selection of the actual data mining model, settling of the testing system, and finally training / assessing of the model.
- Evaluation: As the name of the phase suggests, this phase is the final evaluation of the model and assessment of the same against the set objectives of the project at the beginning.

Question 2. Provide in few sentences the problem statement and the main objective of your project and briefly explain the in bullet points the roadmap (phase one of CRISP-DM)

Solution:

Problem Statement: Prediction of stock prices of Lithium Ion Battery manufacturers using historical data of Lithium Mining Companies.

Explanation: We plan on analyzing the historical data of Lithium Mining companies for prediction. This analysis will include the stock prices of the mining companies and sentiment analysis of reputable financial news outlets' headlines regarding the same. Research of this kind could be beneficial to companies that make heavy use of lithium ion batteries by determining upcoming cost spikes by analysis of the state of the lithium mining companies.

Roadmap:

- Business Objective: We believe that businesses that make heavy use of lithium ion batteries would find our research very profitable.
- Situation Assessment: We plan on focusing on the top lithium mining companies (Albemarle, Ganfeng Lithium, Livent, etc) and using the publicly available historical stock market data for our analysis.
- Data Mining Goal: To find correlation between stock prices of lithium ion battery manufactures and that of lithium mining companies.
- Project Plan: We plan to begin by collecting the data followed by its cleaning. Then, we plan to evaluate between different models and select an appropriate one. Finally, model training and testing would follow ending with the evaluation of the results.

Question 3. Explain the main difference between supervised learning and unsupervised learning.

Solution:

The main difference between supervised and unsupervised learning is that supervised learning deals with datasets that are labeled and the model tries to classify or predict on the basis of the labels it has seen in the training data, whereas in unsupervised learning the training data is unlabeled and the model is trained to analyze the data to find out the hidden patterns (or clusters) in the same.

Question 4. Explain the main difference between supervised learning and unsupervised learning.

Solution:

The differences are pretty much self explanatory. Keyword-based search finds results by searching for the keyword(s) provided by the user. Semantic-based search would not only search for the keyword(s) provided, but also for semantically equivalent keywords. Contextual Search, on the other hand, also considers the context of the user (on the basis of location, time, other searches etc) when providing results. For example, a search for time would yield the current time on the basis of the location (and time zone) of the user.

Question 5. Define the phenomenon of Big Data then one example of each of the following data types:

Solution:

Big Data is characterized by the 3Vs. That is, volume, variety and velocity. Volume, as the term suggests, says that Big Data has huge volume which makes it necessary for the field of data mining to exist. Variety means that the data can be in various formats such as text, excel, pdfs, images, videos etc. Velocity refers to the rate at which data is generated and, characteristically, Big Data grows at a very rapid rate.

- Structured Data: Structured Data refers to well organized data. Data that is in a ready to use format and can be easily analyzed by a human or a computer program. For example: excel sheets with well labeled columns.
- Unstructured Data: Unstructured data refers to data that is unlabeled, uncategorized and generally hard to use in large quantities. For example: various unlabeled or misnamed documents, images, videos in scattered computers and servers.
- Semi-structured Data: Semi-structured data is data that has some elements of organized structure which makes it easier to analyze than unstructured data. For example: labeled zip files with relevant compressed files inside.

Question 6. When you are solving a data science problem, you can map it into one or a combination of one the following data mining tasks described in lecture 2, briefly explain them in no more than 3 sentence:

- **Data Clustering**
- **Data Classification**
- **Building Recommendation Systems**
- **Mining Association Rules**

Solution:

- Data Clustering: Data Clustering is a data mining technique to group or "cluster" data items of similar types. For example: kNN clustering algorithm
- Data Classification: Data Classification is used to classify data into predefined categories or "classes" on the basis of labeled training data given that has been used to train the classifier.
- Recommendation Systems: Recommendations systems use various methods and algorithms to analyze data available to them to produce suggestions or recommendations for users. For example: Watch Recommendations on Youtube or Facebook.

- Mining Association Rules: These types of machine learning algorithms are used to identify causal relationships and associations between data points. For example, in medical field: symptoms and diseases.

Question 7. In paper "the Earthquake Shakes Twitter Users", the author used tweets to build a "classifier" to tell whether a tweet is about an earthquake or not. Briefly list the features used by the author to classify a tweet to relevant or irrelevant to an earthquake event:

Solution:

1. Keywords in a tweet
2. Number of words
3. Context of target-event word

Question 8. Summarize at a high-level data science approach that was used in the work of "Twitter mood predicts the stock market" – Feel free to use diagrams.

Solution:

Text content of Twitter feeds from February 28, 2008 to December 19, 2008 was used for sentiment analysis using OpinionFinder and Google-Profile of Mood States (GPOMS) to try and predict / correlate mood changes with changes in Dow Jones Industrial Average (DJIA).

For data preparation, stop-words (the, and, are, etc) and punctuation was first removed from the text of twitter feeds and then grouped together by date. Also, since this is sentiment analysis, only tweets that seemed to express emotions were retained with the rest discarded. Following that OpinionFinder and GPOMS were fed the data to generate estimate of public mood. Finally, Granger Causality analysis was used to correlate public mood with changes in DJIA.

GPOMS mood indicators showed better causal relation between the public's "calmness" and DJIA values. OpinionFinder's "positive" / "negative" mood indicator proved ineffectual in showing any causal relationship.

Question 9. In Chapter 2 of the book, the concept of recommendation systems was explained. Highlight the difference between "user-based collaborative filtering" and "item based collaborative filtering"

Solution:

User-based collaborative filtering recommendation system finds similarity between users (based on their consumption pattern) and recommends items on the basis of consumption of similar users whereas item-based collaborative filtering finds similarity between *items* and recommends items to users that are similar to already bought/watched/consumed items by the user before.

Question 10. Define the concept of "alternative data" in your own words and briefly explain how it applies to (1) the world of finance and (2) epidemiology (following the COVID-19 Early-Alert assigned reading)

Solution:

Alternative Data refers to non-traditional data such as social media feed, blog posts, web traffic, app usage etc which is used to give companies an edge in the market.

- In regards to finance, sentiment analysis has already shown some promise with regards to stock price assessment and prediction. Various studies have shown causal relationships between social media feed interest in some product and actual profit gained by the company that is the manufacturer of the said product.
- As for epidemiology, the Early-Alert paper shows correlation between people's mobility and COVID-19 case spikes which neatly evinces the use of alternative data in this field.

Question 11. Explain how Google Trends for Flu Prediction went on from success to failure. What led to the failure of predicting flu at later iterations?

Solution:

While Google never revealed exactly how their system worked, it can be surmised that their system was probably prone to overfitting and couldn't distinguish between other seasonal terms and terms that were actually related to flu like symptoms. With the failure it faced at predicting the Swine Flu outbreak, it faced its final nail its coffin and was shut down.

Question 12. In his latest, Bill Gates mentions in Chapter 3: Disease Surveillance. Briefly describe (1) what he meant by Disease surveillance, (2) How can predictive analytics and alternative data sources help in Disease Surveillance?

Solution:

- Disease Surveillance: Gates refers to disease surveillance as the task of watching for disease outbreaks that might prove to be catastrophic and might become epidemics or pandemics. The purpose of the same is, of course, to try mitigate outbreaks before they have opportunity to spread beyond control.
- As Gates mentions, analysis of social media posts and blog posts give a nice supplement to traditional active disease surveillance methods. Predictive Analytics algorithms might prove useful here for the analysis of the above.

Question 13. In the paper "Twitter mood predict the stock market", we discussed how correlation DOES NOT automatically reflect a causation relationship. How did the authors test for causation? Explain the algorithms they adopted.

Solution:

The authors tested for causation by using the Granger Causality analysis to not actually test for causation but instead to test if the twitter mood time series actually had any predictive effect on DJIA values or not.

Question 14. In the Vaccines Hesitancy assigned reading, briefly describe the analytics objective of the work? And provide one limitation of the tool following your own analysis from the reading.

Solution:

The objective was to track and analyze social media vaccine sentiment to better prepare health professionals for vaccination conversations and campaigns.

Limitation: One limitation, I believe, is that this study couldn't actually tell the root cause behind any excessive negative sentiment behind vaccines and just warn people about the same.