BIG Data mining process, problem domain and its techniques with applications

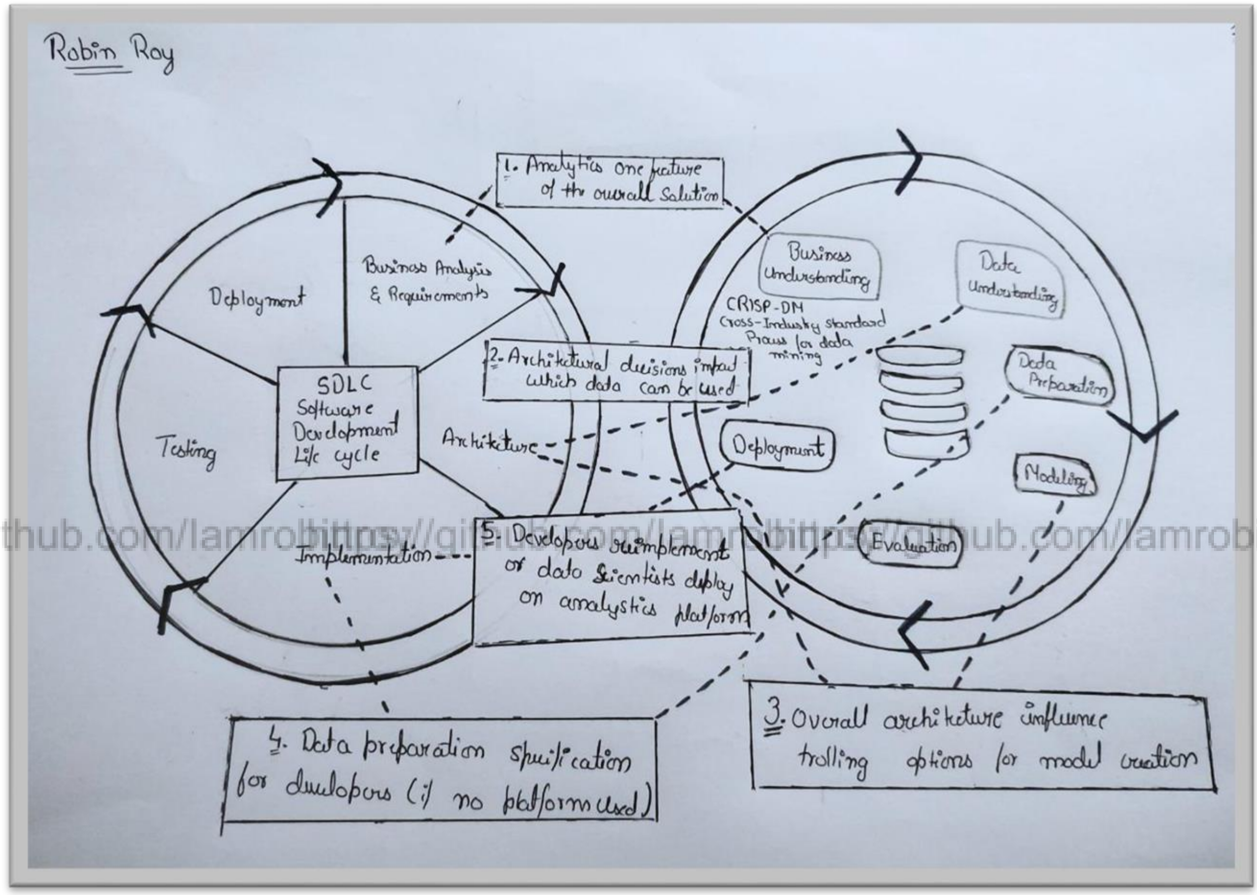
Study by:
Robin Roy

PART A | Data mining methodology

The research paper CRISP-DM in other words Cross-Industry Standard Process for Data Mining is an open source data mining model, developed in collaboration with industry leaders, several tools and using the inputs from hundreds of users using data mining. The data model CRISP-DM was found in late 1996 by Daimler-Benz, Integral Solutions Ltd., NCR and OHRA who are knows as the leader of nascent data mining market. It is always in development of a standard model which supports the community of data miners. The model is being developed as newer extension and improvements are expected.

The model proposes better actions plans that can be followed by organizations in order to reduce the time consumed in data mining and get faster results from it giving better outputs. The basic concept of the CRISP-DM is dividing the process into different stages of action, they are **[1]** a) business understanding, (b) data understanding,(c) data preparation,(d) modelling,(e) evaluation and (f) deployment. Since the CRISP-DM model has 6 phases it enables organizations to divide their data mining process into different phases and gave them structure while starting a project. CRISP-DM gives us a road map to show to start and finish a data mining project which is of very helpful for data mining community whether the person is an expert of just a beginner the CRISP-DM is of great help as already mentioned there are six phases to guide us in the project throughout different processes.

Business understanding is sometimes the underestimated phase, it mainly about what we want to achieve from the model with a perspective of business. Setting objectives means explaining the elementary objective from the business. We need to access the current situation which involves detailed finding of facts about the resources, constrains, assumptions and other things that we need to consider where setting the analysis goals with the project plan which includes setting the resources, knowing the requirements, accessing the risks, and the costs of the project. The phase two is data understanding where the data is acquired and listing the sources from where the data is acquired. The data should be described and explored verifying the data quality. Data preparation, in this stage we decide on what data we are going to use for the analysis. The selection of the data is based on many features like data mining goals the quality of the data and technical constraints like limits on data volume or the types of data. There is also requirement for data cleaning which involves increasing the quality of the data, by checking missing data followed by construction of required data. There is construction of derived attributes and records. The last step is checking the integrity of the data which are being collected from different sources like databases, tables or records to create new records or values. The fourth phase is regarding selecting the specific modelling technique like decision-tree building with C5.0 or neural network generation with back propagation. The result need to be summarized by assessing the model, revising the parameter settings and tuning them for next modelling run. Continuing this process until the best model is found. The fifth stage is evaluation of results where the degree of correctness of the model to achieve business reason and check for the defectiveness of the model. After the model is assessed with the business success criteria the generated models becomes the approved model. The last phase of the CRISP-DM is deployment where the evaluation results are taken to determine the strategy for deployment. The predictive analysis in this step can help the business in this stage. It should be monitored and a maintenance plan should be made. The report consists of the summary of the project and the experiences. The project is reviewed with what went and right and what went wrong and the experience documentation is made including the pitfalls and the best suited techniques for the project.


Robin Roy
1. Analytics one feature of the overall Solution
Business Analysis & Requirements
Deployment
SDLC Software Development Life cycle
Testing
Architecture
Implementation
2. Architectural decisions impact which data can be used
Business Understanding
Data Understanding
CRISP-DM Cross-Industry standard Process for data mining
Data Preparation
Deployment
Modeling
Evaluation
5. Developers reimplement or data Scientists deploy on analystics platform
4. Data preparation specification for developers (if no platform used)
3. Overall architecture influence trolling options for model creation

[2]The paper deals with progression analysis of signals which is like an extension of CRISP-DM to stream Analytics by Pankush Kalgotra and Ramesh Sharda of Oklahoma State University. The paper focuses on the analysis of different types of signals generated considering the time factor. There are different patterns in the signals as observed which represents certain outcomes. The main intension of the paper was to improve CRISP-DM process to insert data preparation processes for the purpose of sequential mining. To illustrate this data of patients with Tobacco disorder is used to develop certain other diseases over multiple hospital visits to find generalizability of progression analysis. Streams are an interesting area of study in Big Data process methodology. The traditional Cross-Industry Standard Process for Data Mining or CRISP-DM to analyse time-stamped data streams, in addition to the six phases of CRISP –DM the improvement will extend its applicability to several domains. It is not the case that the current six phases of the CRISP-DM is not applicable. The phase of data preparation and modelling of streaming data is different from the usual one. The paper brings an algorithm where the data is in form of multidimensional data streams in a form to do sequence analysis, it is useful to use with time-ordered data, and the different patterns are extracted according to predefined minimum support. The paper deals with data which is multi-dimensional and time variant (MDTV). A sample hypothetical system is developed to develop a process to make the data streams and identify the patters in MDTV with a method to analyse data streams. The process of sessionizing the signals is done. nPath function of Teradata Aster, Big Data platform can be used. The progression analysis is very useful in understanding how different diseases arise one by one. In this research paper it is discussed mainly regarding the data preparation and analysis phase of CRISP-DM, where the model can be applied to situations where multiple signals are created with track on time.

[3]Synthesizing CRISP-DM and Quality Management which is a data mining approach used for production process is a research done by Franziska Schäfer, Christian Zeiselmair, Jonas Becker, Heiner Otten of University Erlangen, Germany deals with QM-CRISP-DM for data mining applications with respect to production and improvement of different processes. It suggests better tools for each six phase of CRISP-DM. The QM-CRISP-DM provides quality management tools for the phases. It is a proved method for predicting errors in the system in electronics production domain. Reaching 2020 business began to exploit the potential of data mining while suing the techniques in analysis of their production analysis and forecasting. Companies prefer to use CRISP-DM since it has stepwise procedures. But it is visible that some of quality management techniques are missing in CRISP methodology. In consecutive years the DM arose, starting from the Knowledge Discovery in Databases approach. SEMMA is a type of KDD. Many are using CRISP-DM currently but it is missing connection management methods for the long term run of the generated outputs. The CRISP-DM was later refined with ASUM-DM which contains better structure, management, and templates but it is just restricted to be used with IBM. Development like CASP-DM just gave importance to distributed big data and processes. But an integration of quality management way of thinking has never appeared before discussed in this research.

Artificial Neural Networks to forecast the optimal life span of a machine. They also use logistic regression to reduce the input factors. The business understanding phased is enchased with quality management. The issues and motives of the project are assessed. Tools like Stakeholder/VoC analysis are used to measure quantities in the understanding phase the QM-CRISP-DM can shed some light by giving focus to six sigma to formulate statistical test structure with less parameters in algorithms. In data preparation process limits that were used in after steps of evaluation were separated from utilizing an appropriate I-control chart. In modelling the ANN is implemented utilizing python and TensorFlow framework and Keras library. All the parameters were selected by trial and error method for the purpose of reducing the MSE. The evaluation is improved by benchmarking or internal and external conditions of the company. Deployment and control in QM-CRISP-DM contains documentation of the project and the error predicting algorithm. The QM-CRISP-DM cycle in this paper extended the traditional CRISP-DM six phases with quality management tools. The six sigma and QM-tools gives a strong set of techniques.

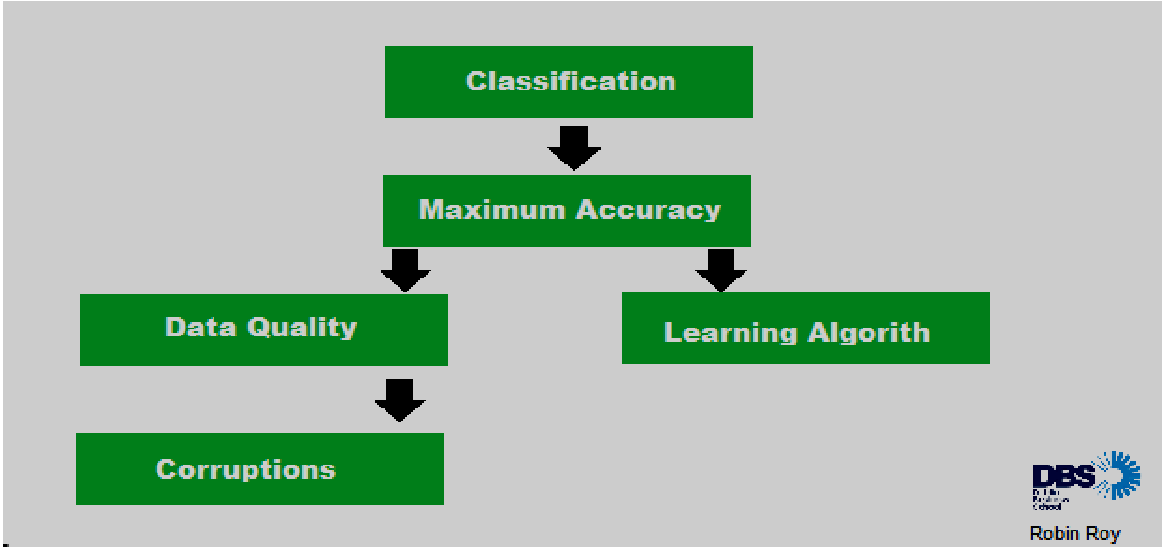
[4]The paper which discuss framework for automating the pre-processing of sensor data analysis by Hiroko Ngashima, Yuka Kato of Tokyo Woman's Christian University, Japan discuss about research based on the CRISP methodology which deals with pre-processing of data to get better outputs, which consists of the process of finding the outliers, understanding missing data, data arranging, integration and normalization. It is important that that CRISP-DM to be improved with pre-processing of data since deleting ambiguities and irregularities from data can improve the model drastically. The proposed framework is APREP-DM which is automates the process of pre-processing of data for data mining APREP-DM is proved to have efficiency in analysing sensor data. The tasks are automated and pre-processing is important in this scenario since there will be several missing of data when it comes to the data outputted by sensors, there is no proper examination of pre-processing of data to correct ambiguities in CRISP-DM as it does not treat outliers in the phase III. But still CRISP-DM is used with APREP-DM where it do not deal with specific products. CRISP-DM do not have any outlier detection system in pre-processing. The frameworks APREP-DM SEMMA, KDD, have this feature. KDD has the ability to delete outliers on the other hand SEMMA and APREP-DM finds the existing outliers. KDD has two steps and is most compact. The APREP-DM adds additional automated pre-processing step to the CRISP-DM.

CRISP-DM is evolving throughout the years since new technologies require methodologies that suit its faster development and completion. The above discussed research papers and journals exhibit the proof of continuous evolution of CRISP-DM. The frameworks and methodologies provides with proven architecture in the BIG Data mining problems. Coming to the year 2020 we can see the CRISP-DM already made lot of improvements through the evolving researches.

PART B |Data mining problem domain and applications

Innovation transformation has been encouraging a large number of companies to produce huge information like in sensor data that create nonstop surges of computerized information, bringing about what has been called as "Big Data". This writing focuses on analysing, synthesizing and presenting a comprehensive analysis of 'Big Data mining problem'. It has been an affirmed wonder that tremendous measures of data have been large consistently produced at remarkable and regularly expanding scales. Big data sure includes an incredible assortment of data structures: text, pictures, recordings, sounds, and whatever, and their discretionary mixes. Big data habitually comes as floods of an assortment of types. Time is a vital dimension of data streams, which frequently suggests that the data must be handled or mined in a constant way.

Additionally, the current significant users of big data, companies, are particularly keen on a big data condition that can quicken an opportunity to-address basic business addresses that exhibit business esteems. In the data mining point of view, mining big data show numerous new difficulties and challenges. Despite the fact that big data possess very good usefulness since it has concealed information and increasingly important bits of knowledge within them. It carries huge difficulties to extricate these concealed information and insights from big data since the built up procedure of information finding and data mining from traditional datasets was not intended to and won't function admirably with big data. The challenge is what to mine this enormous volume of data with the goal that it can be investigated by a data mining model. The exhibition, which we generally need to expand, of the classifiers worked under such conditions will vigorously rely upon the nature of the training of the data, yet in addition it depends on the validity of the classifier.



The existence of noise in the data may influence the natural qualities of a characterization issue, since these deformities could present new properties in the issue domain. Noise can lead to forming shorted clusters of a specific class in parts of the domain related to a different class. The limits of the classes and the overlaps between them are additionally factors that can be influenced as an outcome of noise. Every one of these adjustments troublesome the information extraction from the data and ruin the models acquired utilizing the noise data when they are contrasted with the models gained from clean data, which speak to the genuine verifiable information on the issue. Large number segments decide the nature and quality of a dataset (R.Y. Wang, V.C. Storey, C.P. Firth, A Framework for Analysis of Data Quality Research, IEEE Transactions on Knowledge and Data Engineering). Noise happens when a model is mistakenly named. Noise can be ascribed to different causes, for example, subjectivity during process of naming, errors during the insertion of data, or insufficiency of the data. Data is stored in variety of systems and exists in multiple formats. **Distributed databases** are one of the problems in Big Data mining domain. **[5]** Architecture made of data grid generic, and specific data mining grid services that dynamically configured to the application's data mining requirement. Makes the data mining process easy in an environment which shows characteristics of a distributed database. The layers of the mentioned grid infrastructure are arranged by its heterogeneity and it does not affect the grid. Grids provide access to conveyed processing and data assets, permitting data-serious applications to improve vastly in accessing data and the way its managed and analysed.

[6] In the process of mining **complex data** objects in other words it can be describes as generalization of structured data comprises of setting a valued attribute where the process is of generalizing every value in the list to its respective higher-level concepts. The number of elements in the list, the types of elements, the range of those elements, or its average, this is in case of set-valued attribute. In list valued- attributes which is similar to set-valued attribute except the order of elements in the sequence must be observed in the generalized. Data size and configuration is not in a particular format so it's hard to keep up **security and privacy** of one client from another. The quantity of algorithm dealing with security is not limited. At the point as the size of data alters or arrangement differs there is need to apply other algorithms. We characterize the security or protection calculations to it can't be appropriate to overhauled data. For example In emergency clinic the data gathered and it might redesign day by day and it might be in various configuration, so it gets hard to examine and make sure about the recently included data. As data is connected with such a large number of formats and it is difficult to keep security of data and consequently and is a huge problem in data mining. The **speed and velocity** with which data is produced is also an area of discussion in the field of data mining. Since data is created in a very high velocity there no doubt that separate strategies are required for its processing. A scenario where data that is created with high speed and velocity would be Twitter tweets or Facebook posts. Speed alludes to novel speed is timely way. In many scenarios it is hard to keep up novel speed on account of difference in the type of data and quantity of data.

Data is changing **healthcare services** making better patient results while lessening the treatment costs. The healthcare department suppliers started exploit the potential of immense amount of data and started making them stand out. Every patient can be treated with custom made treatment protocols that will vastly benefit the patient's recovery. Genetic algorithm includes three stages determination, hybrid and transformation for each gene. Genetic algorithm has discovered approach in phylogenetic, computational science, engineering, financial aspects and much more. The algorithm can be utilized as a classifier in numerous domains as it can be used for the purpose of prediction and analysis. K-Nearest Neighbour (KNN) and Rule Induction are techniques used in healthcare genetic algorithm.

[7] Big data creates the link between patients, doctors, diagnosis and predicting the surgery and pharmaceutical companies. Also Big data haves the computing power too high to study DNA sequence and predict the disease or required pattern. Big data can also be helpful to monitor premature babies and sick baby unit. By analysing every heartbeat now a days it can be possible to identify the disease before its actual symptoms. It is possible to create connection between patients, the people who treat them, and the diagnosis. It is helpful in prediction for testing and pharmaceuticals. The figuring power is too high to even consider decoding DNA succession and find expected diseases. Benefits are proven in monitoring premature and sick infants. By analysing heartbeat nowadays it is very easy to distinguish the disease before any proper manifestations. In general ANN is called as "Neural Network". ANN is a non-linear statistical data modelling approach and used to manage complex relationships between I/P and O/P. As dataset used in ANN grows massively we need to analyse it automatically. It is also helpful to recognize the pattern from which it belongs. Classification, Prediction, Clustering & Association Rules are the steps of data mining and are useful in neural network to identify patterns. Recognize indications of extortion squander and misuse (FWA) early and take preventive measures. Enhance tasks to convey the most ideal consideration at the least expense. Increase persistent understanding and innovative treatment models. There is Improvement in safety and capacity to foresee medicine fraud by distinguishing them learning the patterns.

REFERENCES

[1] Colin Shearer The CRISP-DM Model: The New Blueprint for Data Mining

[2] Pankush Kalgotra, Ramesh Sharda (2016). Progression Analysis of Signals: Extending CRISPDM to Stream Analytics. 2016 IEEE International Conference on Big Data (Big Data)

[3] Franziska Schäfer, Christian Zeiselmair, Jonas Becker, Heiner Otten(2018). Synthesizing CRISP-DM and Quality Management. A Data Mining Approach for Production Processes. 2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)

[4] Hiroko Nagashima, Yuka Kato APREP-DM(2019). A Framework for Automating the Pre-Processing of a Sensor Data Analysis based on CRISP-DM. 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)

[5] Alberto Sanchez, Pilar HerreroFacultad (2014). Improving Distributed Data Mining Techniques by Means of a Grid Infrastructure. Universidad Polit´ecnica de Madrid, Madrid, Spain.

[6] Jiawei Han, Micheline Kamber, Jian Pei (2012). Data Mining Techniques and concepts. Data mining third edition

[7] Asha M Pawar (2015) Big Data Mining: Challenges, Technologies, Tools and Applications. 2015 SKNCO.