Advanced Network and Data Mining: CA 1

Submitted By: Vaibhav Tiwari
Student Number: 10621051
Lecturer: Terri Hoare

Part A:

Read the journal article available on Moodle "The CRISP-DM Model: The New Blueprint for Data Mining" Shearer 2000. Write a critique of this article as it applies to the mining of 'Big Data' in 2021. Your appraisal should include a review of two related journal articles.

Introduction:

The report includes a critique of the article "The CRISP-DM Model: The New Blueprint for Data Mining" Shearer 2000 as it applies to the mining of 'Big Data' in 2021. I have also included a review of two related journal articles to support the usage of CRISP-DM in big data mining and suggested improvements and changes required for it to be more scalable with respect to big data. This report emphasizes how CRISP-DM model, which is more than two decades old, continues to be the backbone for Data mining today and how changes are required to this model to make it scalable and practical for big data mining. We have highlighted the strengths and weaknesses of the model and used two reference papers to demonstrate them.

CRISP-DM: Critique and review based on two reference papers:

Colin Shearer's (2000) article, "The CRISP-DM Model: The New Blueprint for Data Mining", which was published over two decades ago provides a detailed overview of the CRISP-DM model, a popular model for data mining. Since then, there have been many changes with respect to the introduction of big data. One of the major criticisms of the CRISP-DM model is that it was designed for smaller, well-structured data sets, and it may not be as effective for dealing with the massive, unstructured data sets that are becoming increasingly common in the age of big data. The model's emphasis on the data preparation stage can also be problematic in big data scenarios, as it can be time-consuming and resource intensive. Another limitation of the CRISP-DM model is that it assumes a linear, sequential process, which can be difficult to follow in big data scenarios where new data is constantly being generated, and analysis needs to be done in real-time. Additionally, the model does not place enough emphasis on the importance of visualization and communication, which are critical for effectively communicating insights and findings to stakeholders.

One of the most widely used frameworks for conducting data mining and analytics projects is the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. As of today, the CRISP-DM methodology is still relevant and widely used for mining data. However, a few changes/modifications are required to adapt this to big data projects due to the increase in data size and complexity. The methodology consists of six phases - Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment. With respect to the big data in the data understanding phase, it is important to consider the use of big data technologies such as Hadoop or Spark to handle large volumes of data.
In the data preparation phase, data cleaning and feature engineering may be more challenging due to the large, diverse data sources. Thus, we need automated tools or machine learning algorithms to assist with this process. In the modelling phase, the traditional algorithms are not scalable to big data. Hence, we require distributed or parallel algorithms. During the evaluation phase, it is important to consider the metrics for big data models, such as processing time, accuracy and memory usage which are calculated based on the traditional and performance metrics.

These stages have been explained further below -

Business Understanding: The first stage of the CRISP-DM methodology involves understanding the business problem and the objectives of the project. In big data projects, it is important to have a clear understanding of the specific business requirements and how they relate to the available data sources. It is also important to identify the key performance indicators (KPIs) that will be used to measure the success of the project.

Data Understanding: The next stage of the CRISP-DM methodology involves understanding the available data sources and their quality. In big data projects, this can be particularly challenging due to the volume, variety, and velocity of data. It may be necessary to use specialized tools and techniques to gather and process the data, such as Hadoop, Spark, or other distributed computing platforms.

Data Preparation: Once the data sources have been identified and assessed, the next stage involves preparing the data for analysis. This includes cleaning the data, transforming it into a usable format, and creating any derived variables or features that will be used in the analysis.

Modeling: In the modeling stage, various machine learning techniques can be applied to the prepared data to develop predictive models or other analytical solutions. In big data projects, this may require the use of distributed machine learning platforms, such as TensorFlow or PyTorch.

Evaluation: The evaluation stage involves testing the performance of the models or analytical solutions and validating their accuracy and reliability. In big data projects, this may require the use of specialized validation and testing tools to handle the large volume of data.

Deployment: Once the models or analytical solutions have been validated, they can be deployed for use in the business environment. In big data projects, this may require the use of distributed computing platforms or cloud-based services to handle the volume and complexity of the data.

Overall, while the CRISP-DM methodology provides a solid foundation for data mining and analytics projects, modifications may be necessary to effectively apply it to big data mining projects. Users need more flexible, agile, and iterative approaches to data mining. While the principles of CRISP-DM model are still relevant, a lot of updates need to be considered due to the challenges presented by big data.

Based on the reference paper - Reza; Khan et al. (May/Jun2018, Vol. 8 Issue 3, p1-1, 14p) the context of mining big data in 2021, the CRISP-DM model remains relevant and valuable for several reasons:

Clear guidance: The CRISP-DM model provides a clear and structured roadmap for approaching data mining projects, which can be particularly useful when dealing with large and complex data sets.

Flexibility: While the model follows a structured process, it allows for flexibility to adjust and modify each stage according to the specific needs of the project.

Collaboration: The model encourages collaboration among different stakeholders involved in the project, including business analysts, data scientists, and domain experts.

Iterative approach: The CRISP-DM model recognizes that data mining is an iterative process, and it provides a framework for continuous improvement and refinement of the models.

Risk mitigation: The model promotes a proactive approach to risk management, including identifying potential risks and addressing them early in the process.

Overall, the CRISP-DM model remains an important framework for guiding data mining projects in the era of big data, providing clear guidance, flexibility, collaboration, and risk mitigation.

The paper - "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories" by: Martinez-Plumed, F.; Contreras-Ochando, L.; Ferri, C.; Hernandez-Orallo, J.; Kull, M.; Lachiche, N.; Ramirez-Quintana, M.J.; Flach, P.. In: IEEE Transactions on Knowledge and Data Engineering IEEE Trans. Knowl. Data Eng. Knowledge and Data Engineering, IEEE Transactions on. 33(8):3048-3061 Aug, 2021; USA: IEEE Language: English, Database: IEEE Xplore Digital Library, discusses the evolution of the Cross Industry Standard Process for Data Mining (CRISP-DM) framework over the past 20 years and its current relevance in the context of modern data science practices.
Martinez-Plumed, F. et al. (2021) provide an overview of the CRISP-DM methodology and how it has been used in various data mining projects. They also discuss how the framework has evolved over time to incorporate new techniques and technologies, such as big data and machine learning.
The paper highlights the importance of a flexible and iterative approach to data science projects, and how the CRISP-DM framework can be adapted to suit different project needs. The authors also discuss the challenges and opportunities presented by the increasing complexity of data science projects, such as the need for interdisciplinary collaboration and the importance of ethical considerations.
Overall, the paper emphasizes the continued relevance of the CRISP-DM framework in guiding and structuring data science projects, while also acknowledging the need for ongoing adaptation and evolution in response to new challenges and opportunities.

Conclusion:

Based on the journals reviewed, we can conclude that CRISP-DM model is still the base selection for Data mining of big data. However, there are various additional and scalable enhancements required in each stage of the model which make it better and more efficient.
The major changes are required while preparing the data and for preprocessing the huge datasets for higher accuracy and efficiency in results. The model is more than two decades old but highly relevant to all the current implementations. The choice of methodology will depend on the specific needs of the project and the resources available to the team.

References:

1. Colin Shearer, 2000, "The CRISP-DM Model: The New Blueprint for Data Mining", provided by Teradata university network.

2. Martinez-Plumed, F.; Contreras-Ochando, L.; Ferri, C.; Hernandez-Orallo, J.; Kull, M.; Lachiche, N.; Ramirez-Quintana, M.J.; Flach, P.. (2021), "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories", In: IEEE Transactions on Knowledge and Data Engineering IEEE Trans. Knowl. Data Eng. Knowledge and Data Engineering, IEEE Transactions on. 33(8):3048-3061 Aug, 2021; USA: IEEE Language: English, Database: IEEE Xplore Digital Library

3. Mohd Selamat, Siti Aishah; Prakoonwit, Simant; Sahandi, Reza; Khan, Wajid; Ramachandran, Manoharan (2018),"Big data analytics—A review of data-mining models for small and medium enterprises in the transportation sector";WIREs: Data Mining & Knowledge Discovery , May/Jun2018, Vol. 8 Issue 3, p1-1, 14p; DOI: 10.1002/widm.1238, Database: Library & Information Science Source

Part B:

Select a Big Data mining case study published either in a journal, conference paper or vendor report. Discuss the data mining techniques applied and tools used. Highlight the benefits to the business together with measurable implementation success criteria.

Introduction:

As a part of this report, I have analyzed a case study of the online customer reviews and satisfaction of guests in Hotel Atlantis collected over 3 years between 29th October 2018 and 29th October 2019. I have highlighted the methods that have been used in the process of collection, processing and analysis of the data collected over the duration mentioned above. In the article by Shengnan Wei et. Al. (2022, Page 4), the following hypotheses were indicated as a part of the big data analysis:

“Hypotheses 1(H1) – Service has a positive impact on customer satisfaction at Atlantis, The Palm.

Hypothesis 2(H2) – Water park facilities have a positive impact on customer satisfaction at Atlantis, The Palm.

Hypothesis 3(H3) – Hotel star ratings have a positive impact on customer satisfaction at Atlantis, The Palm.”

The study was conducted around the customer service reviews provided by the guests and aimed to analyze and review the impact of Tourism and facilities provided at hotels to understand customer satisfaction and to provide insights on how to upscale hotel managers when formulating and implementing strategies and tactics to improve customer satisfaction.

Case study - Methodology and Implementation:

The aim of this research was to examine customer satisfaction with Atlantis, The Palm by analysing online reviews from the Google Travel webpage. To analyse the reviews, the researchers used a text mining method to extract keywords, followed by calculating the frequency of keywords and analysing the degree and eigenvector centrality of the top 50 words. CONCOR analysis was then carried out to group the keywords, resulting in four groups named

'Service', 'Dining', 'Scenery', and 'Facility'. An exploratory factor analysis and linear regression analysis were also conducted, which reduced the original 50 keywords to 16 and grouped them into four factors named 'Facility', 'Value', 'Dining', and 'Service'.
The article reviewed the customer service reviews provided by the guests. A total of 2051 reviews were collected using SCTM 3.0 (Smart Crawling and Text Mining) which is a platform developed by Wellness & Tourist big Data Institute. The aim was to collect keywords from the reviews and apply an in-depth analysis on customer reviews. These reviews were used to understand the positive and negative influences on overall customer satisfaction. Shengnan Wei et. Al. were able to analyze that the hotel that had theme parks attracted more tourists. "The Palm Hotel built a popular water park and got listed as one of the top tourist destinations in Dubai", based on the TripAdvisor website.
Dubai aims to reduce its dependency on the primary source which is Oil for generating more revenue and boost its domestic tourism market. Since Dubai has shifted its economic focus, it must also adapt and follow certain standards in the tourism industry. This study studied factors that were closely related to customer satisfaction levels. Big data analysis – Sentiment analysis was used to assess this data and determine the business performance. It was helpful to obtain an idea of the true feeling of the customers through the management.
R language, RStudio and UCINET software were used to conduct sentiment analysis on the data collected using SCTM 3.0; To assess these factors and to collect a meaningful dataset that could be analyzed using the tools, Google travel reviews were collected for 3 years. From the data extracted by SCTM, RStudio was used to extract high-frequency words and matrix. This result was further analyzed on UCINET to determine word centrality and CONCOR. Data was also processed by NetDraw, which created a network diagram for visualization.
In summary the following tools and languages were used in the respective stages –

1. Data Collection – SCTM 3.0, Online reviews from Google Travel
2. Data Processing – Rstudio(Text Mining), UNICET(centrality), NetDraw(Visualization)
3. Data Analysis – Frequency Analysis, CONCOR Analysis, Exploratory Factor Analysis, Linear Regression Analysis

From the top 50 most frequent keywords collected, there were some repeated, unnecessary words and symbols which misled the result. These were hence dropped as they were cleaned and prepared for further processing. The table of top 50 words collected had basic parameters like the Ranking, Word, frequency, and the percentile of the number of times it had appeared in the reviews.
Post data cleaning, the result indicated that the customers chose the destination for its value rather than the basic hotel functions there. (Page 5, Paragraph 1)
Stage 2 of processing the data using UCINET displayed a comparison of keyword frequency and centrality rankings. The words were enlisted under three labels – Frequency ranking, degree ranking and eigenvector ranking. This data was processed using CONCOR analysis which divided the words into four categories – Facility, service, dining, and scenery. The visual results and the tabular data generated by CONCOR related to the perfect facilities which Atlantis, The Palm provides based on customer reviews.
Factor analysis with a minimum factor loading set at 0.400 extracted 16 words from the top 50 words by reducing the large number of variables to small number of variables. This allowed the main variables to come out and represent the total variables, thus covering 48.032% of all variances. The KMO figure generated was accepted as the value was greater than 0.5, hence, factor analysis was suggested suitable for the case study. After factor analysis was performed, a regression analysis of the customer experiences and overall satisfaction was conducted. variable (four independent variables (Facility, Value, Dining, Service) and one dependent

variable(Customer Satisfaction) were used. Following results were drawn –
Overall variance: 3% ($R^2 = 0.03$)
Standard error of the estimated value: 0.873
The correlation between the dependent and independent variables was low which might have been caused by loss of some words due their low frequencies in reviews.
Value and Dining were significant at $p<0.005$ level and both predictors had negative, standardized coefficients which indicated that they were negatively related to customer satisfactions. Thus, leading the study to understand that the Palm hotel needs to pay more attention on these two aspects. The other two factors (Facility and Service) had no relationship with overall customer satisfaction based on the Linear regression analysis result. Shengnan Wei et. Al. thus concluded that the facility had no direct relationship to customer satisfaction. The results thus failed to support the hypotheses stated.

Understanding the results:

The results showed that the factor with the highest beta coefficient was 'Value', containing words such as 'Trip', 'Service', and 'Family', indicating that customers choose Atlantis, The Palm based on their perception of the hotel's special value. The second-highest beta coefficient factor was 'Dining', containing words such as 'Dinner', 'Buffet', and 'Food', suggesting that customers have high expectations of the dining experience. However, the linear regression analysis showed a negative relationship between the 'Value' and 'Dining' factors and overall customer satisfaction. The 'Facility' and 'Service' factors showed no relationship with overall customer satisfaction.
These results suggest that the hotel should focus on improving the value and dining experience to meet customers' expectations. The high frequency and centrality of the 'Facility' keywords did not have a direct relationship with customer satisfaction, which contradicted previous hypotheses. Additionally, the star rating of the hotel had a negative relationship with overall satisfaction.

Conclusion:

The objective of this research was to investigate customer reviews of Atlantis, The Palm to identify key attributes of customer satisfaction. The results showed that customers frequently used positive words such as 'amazing' and 'great', and words related to water and parks were also frequently mentioned. However, the study found that overall customer satisfaction did not have a direct relationship with hotel facilities and service, according to exploratory factor analysis and linear regression results. The study applied semantic analysis to a relatively new field and conducted empirical customer satisfaction analysis through big data analysis. The study's significance lies in its potential to provide hotel managers with a better understanding of customer satisfaction attributes, allowing them to formulate functional and valuable managerial strategies.
The study's limitations include the time range of the data collection, which spanned only three years, and the fact that it focused solely on Atlantis, The Palm. Future research could collect data from a longer time range and include comparative research on similar hotels to determine whether water parks contribute to overall customer satisfaction. Additionally, future research could use different analysing tools to investigate customer satisfaction attributes. Despite the study's limitations, the Atlantis, The Palm management team should focus on improving the value and dining aspects of their hotel to increase customer satisfaction.

References:

1. Wei, Shengnan; Kim, Hak-Seon (2022),"Online Customer Reviews and Satisfaction with an Upscale Hotel: A Case Study of Atlantis, The Palm in Dubai";Information (2078-2489) , Mar2022, Vol. 13 Issue 3, p150-N.PAG, 12p; DOI: 10.3390/info13030150, Database: Library & Information Science Source