A Case Study of Evaluating sales of video game with Data Mining Tools and CRISP-DM Methodology

Stage1

Yingzhang Liu 440517742

Business Understanding

o Determine Business Objectives

♣ Background

GameShop is a retailer that sales video games. It is a small and new company which is located in Sydney and was established in 2016 by Jack Jackson. Jack was studying Bachelor of Engineering in the Empire State University. Jack loves video games so much, he was always dreaming about selling video games in the future. After his graduation, he got a huge investment from his father and then he started his company.

In this company, Jack hires four people as salesman, and their job is sell video games in the physical store. He also hires a manager who should in charge of employees' daily work performance and evaluation, customers' complaint and he can make decision in some urgent situation. And also, there are two accountants who record daily income, daily purchase. For Jack, he purchases video games.

Jack is new in this industry and he has no experience doing this. In his company, there is no professional analysis group and he doesn't have analysis skills to help him do his job. He doesn't know how many number for each game he should purchase. So for every hot video games, he purchases 200 for each of them. But he didn't really do the research for these hot games. He thought those hot games are popular because he saw their post and advertisement everywhere. He doesn't spend time on the sale statistics of each game. For other games, which are not "hot", he purchases 100 for each of them every time.

After several months, when Jack check the monthly record about the income. He found that the income is much less than expense. And also, there are a lot of "hot" games that are still stay in the warehouse.

To save his own business, Jack started to look for the solution for his problem. He thought the problem might be the poor management of the manager and the poor skills of the salesmen. He fired one of the salesmen and asked the manager to put more advertisements and sale promotion. It became a little bit better, but when Jack purchased new games again, the problem came out again.

The current solution could attractive more customer to come, but at the meantime it also decreases the profit that Jack could get. It helps sell more games, but every time after Jack purchase new game, a lot of games pile up in the warehouse and it is not easy to sell all of them.

♣ Business Objectives

Jack wants to make the profit reach the maximum, he doesn't think sale promotion would be a long-term strategy and at the same time, Jack would like to know how to purchase game every time, in other words, he wants to find out which game he should purchase more, which game is easy to sell, so that later there wouldn't be much game left in the warehouse.

♣ Business Success Criteria

Objective:

- Get more profit from selling the games.
- Purchase should be more reasonable. In other words, the amount of purchase for each game should be different according to its popular level.

Subjective:

- In the next season's evaluation report, the number of profit should grow up instead of getting lower.
- The number of remaining games should be in a reasonable range.
- The number of less popular game should not be more than the number of popular game.

o Assess Situation

This is the first time that Jack attempt at data mining, he decided to consult a data mining specialist to help him get started and try to analyze data to know different sale statistic for each game in genre, published year, published area and so on.

♣ Inventory of Resources

Hardware: A computer that can access the sales data of every video game using excel.

Data: The data is about the number of sales of video games. The number of sales include sales in north America, Europe, Japanese and other places. The data is public and can be accessed by a website called Kaggle. People have to sign up to access the data.

Personnel: Since Jack just started his small business and he has no experience doing data mining before, there is no any data analysis group or department in

his company. But Jack hired a data mining specialist. If this data mining technique is helpful, Jack will consider to have specialized data group.

♣ Requirements, Assumptions, and Constraints

Requirements:

- The data is generated by vgchartz.com, and the research result of the data should be securely kept in company's computer and only be accessed by Jack and the data mining specialist.
- The data mining specialist should submit the final report about the data within 14 days.
- The report should show the top genre by revenue in current year.
- The report should show the best-selling games in current year.
- The report should show the top publisher by revenue in current year
- The report should show the top platform by revenue in current year

Assumptions:

- Jack couldn't afford the price to hire the best data mining specialist, the one he hired might not provide the correct result that he wants.
- The data is found on website, the quality of data might not be 100% accurate.
- Jack want to simply view the results, and he expects the specialist could explain any diagram, which is generated by the data, to him.

Verify Constraints:

- The dataset is public resource, it is free to access, and could be used for private purpose, but need sign up in the website first.
- The budget just cover the salary for the specialist. There is no budget for purchasing data mining tools. They are just using excel. Other tools might be needed and purchased in the future.

♣ Risks and Contingencies

Risk	Contingency Plan
The specialist might be absent or the work might get delayed.	Hire one more specialist.
The data quality could not be guaranteed then the result could not be completely accurate.	Check the data before getting started.
Even the result would come out, there is still some chance that result would not solve Jack's problem.	Have to find problems in other aspect for example, marketing strategy.

The process could be time consuming.	Pay extra money for overtime work or hire more workers.

♣ Terminology

Data Mining: The practice of examining large pre-existing databases in order to generate new information.

CRISP-DM: A freely available model that has become the leading methodology in data mining. It provides guidelines for organized and transparent execution of any project.

♣ Costs and Benefits

The data is free to collected and the cost would be salary for the specialist. Also, some other tools might be needed, that could be considered as potential cost. After the data mining done, Jack would know how to purchase video games each time. And that would increase the profit for his company and he would not deal with a large amount of video games each time.

o Determine Data Mining Goals

♣ Data Mining Goals

Use historical data to generate diagram for:

- The top genre by revenue in current year.
- The best-selling games in current year.
- The top publisher by revenue in current year
- The top platform by revenue in current year

♣ Data Mining Success Criteria

- The diagram should be easy to understand by Jack.
- The data mining would provide specific number.
- The different between each category should be clearly to see.

o Produce Project Plan

♣ Project Plan

Phase	Time	Resources	Risks
Business understanding	1 day	Data mining specialist	Economic change
Data understanding	3 days	Data mining specialist	Data quality problems, technology problems

Data preparation	5 days	Data mining specialist	Data quality problems, technology problems
Modeling	2 days	Data mining specialist	Technology problems, inability to find adequate model
Evaluation	2 day	Data mining specialist	Economic change, inability to implement results
Deployment	1 day	Data mining specialist	Economic change, inability to implement results

♣ Initial Assessment of Tools and Technique

Use excel to understand and analyze the data at the beginning step.

Data Understanding

o Collect Initial Data

♣ Initial Data Collection Report

The data they use is a public resource on Kaggle website. Once sign up on the website, the dataset is free to download and use. The dataset contains sale statics for more than 16,500 games. It has detail information like rank, game name, genre, publish year, platform, publisher and number of sale in different area, these information could be really helpful because it could compare the number of sale in many ways. At the current situation, Jack has no plan to purchase extra datasets, he only provides the data that was found on Kaggle website, and ask the specialist to use it. However, for further research, if higher requirement is needed they might consider to purchase other valuable datasets which would have more details. For example, different age group must have different interest in different type of game. If Jack could have datasets that cover every area, it would be more accurate to help Jack.

o Describe Data

♣ Data Description Report

Data Quantity:

- The data has number and string.

- Rank, year, NA_sales, EU_sales, Jp_sales, Other_sales, Global_sales are numbers.
- Name, platform, genre, and publisher are strings.
- The dataset is downloaded from Kaggle website.
- The dataset has 16,599 rows and 11 columns (with labels).

Data Quality:

- The data has the most important thing we want, which is the number of sales.
- Two data types: numeric and categorical (string).
- The global_sales for each game could tell whether the game is popular or not.
- The name and sales would be the highest priority attribute.
- The platform, genre, publisher would be lower priority attribute
- The year would be the least priority attribute