# Advancing Vollie's Predictive Capability

MAST90106 Data Science Project Part 1
Semester 1 2020


Group 24

Carlos Davalos Castrillon (1020724)
Sayan Chatterjee (965301)
ZhiZhang Lin (957549)
Shengyi Zhao (990160)


A report submitted in total fulfillment for the
degree of Master of Data Science
under the Faculty of Science

THE UNIVERSITY OF MELBOURNE

May 2020

THE UNIVERSITY OF MELBOURNE

# *Abstract*

MAST90106 Data Science Project Part 1


Carlos Davalos Castrillon (1020724)
Sayan Chatterjee (965301)
ZhiZhang Lin (957549)
Shengyi Zhao (990160)



Vollie is an online marketplace that connects skilled people to non-profits and charities for skills-based online volunteering. Vollie is a start-up company and aims to bring in more charities and volunteers to use the online volunteering platform.

The current platform lacks some capabilities which are critical to enhance user experience to use the platform. Two such capabilities are 1) recommendation of jobs to volunteers based on his skills, interests and work experience, 2) recommendation of a list of volunteers to charities based on matching of project needs and volunteer profiles. Project team had multiple discussions with Matthew Boyd, Chief Executive Officer, Vollie Pty. Ltd. to understand the current business process and problems and agreed that Data Science can help significantly to solve these two problems for Vollie.

Data Science Project Team "Group 24" has analysed the data received from Vollie v1.0 system. The project team has also gone through a number of books, published articles, journals and literature to gain an understanding of the existing research and debates relevant to recommendation systems. Based on the literature review, project team has selected four methods including collaborative filtering recommendation, graph kernel, internal links and link prediction. These methods will be implemented and evaluated, and the best fitted method will be reported for deployment.

This following sections in this document describes the problem statement, insights of Vollie's data, the literature reviews done by the project team, models proposed and a high-level plan including timeline.

# Declaration of Authorship

We certify that this report does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of our knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.

05/06/2020

Sayan Chatterjee (965301)

Carlos Davalos Castrillon (1020724)

ZhiZhang Lin (957549)

Shengyi Zhao (990160)

# *Acknowledgements*

We would like to express our sincere gratitude to our subject coordinator Prof. Michael Kirley, for his patience and guidance throughout the course. We would also like to thank our supervisor Ahmad Asgharian Rezaei, for his continued support.

We want to thank Matthew Boyd, CEO of Vollie Pty. Ltd. for giving his valuable time to help us understand Vollie's business processes, objectives and problems.

Finally, thanks to our family and friends on the three different continents who have helped us in ways unknown to them.

# List of Figures



# List of Abbreviations

**CSR** Corporate Social Responsibility

# Contents

Chapter 1

# Introduction

Vollie Pty. Ltd. is a privately owned organisation founded in the year 2016 which is based in Melbourne, Australia. Vollie projects are exclusively online, meaning that volunteers can donate their skills and experience from anywhere in the world, and around their busy personal and professional schedule. This organisation helps the non-profit organizations to find volunteers with skills they need. It also helps specific businesses to achieve their Corporate Social Responsibility (CSR) objectives. Approximately 2600 volunteers and 400 charities are using Vollie.

## 1.1 Problem Statement

In current Vollie platform, to apply for a job, a volunteer needs to search manually by entering a number of relevant keywords. Therefore, whether a volunteer would be able to find jobs of his interests or skills, entirely depends on selection of keywords used for search. As a result, volunteers are missing out lots of opportunities which are available in Vollie and are relevant to their profiles, but not visible to them. Charities are also not receiving as many applications they would like to have. According to data received from Vollie, approximately 24% of the jobs posted by charities did not receive any single application.

On the other hand, charities manually evaluate candidates who have applied for the job based on their profiles. There is no intelligent capability in the current platform to rank the candidates based on skills, work experience or interests. Charities spend significant time to manually go through the candidate profiles and having telephonic discussions. Due to time constraint, charities are not evaluating all the profiles. According to data received from Vollie, approximately 47% of total applications are in "Pending" status which means that these applications have not even evaluated by the charities.

These issues are resulting in poor user experience and is ultimately leading to loss of business. Data Science Project Team "Group 24" have agreed with Mathew Boyd, Chief Executive Officer of Vollie Pty. Ltd. to develop a machine learning based recommendation system which can significantly enhance the user experience on Vollie's online marketplace.

## 1.2 Project Goal

The primary goal of this project is to develop a "push" scenario via the development of a machine-based algorithm, designed to shortlist best-matching jobs tailored to the volunteer, based on factors such as volunteer data, interests and charitable causes, skills and qualifications, and previous pattern of job applications.

A secondary or "stretch" goal was also established with the client, to develop a machine-based model capable of providing charities with best matching candidates out of Vollie's pool of volunteers, based on factors such as project needs and skill requirements, volunteer profiles, previous interactions, volunteer track record and completion rates.

Hence, the overarching purpose of the project is to improve user experience by applying Data Science.


## 1.3 Challenges

The project team has received data from Vollie in the form of 64 “.csv” files. These files have all data from Vollie related to the entities and fields needed for this project. However, the volume of data is small. For example, only 1875 records related to job applications and Only 1097 records related to projects. Some of the fields relevant to the project have few records. For example, feedback about volunteer is available for only 47 projects out of 1095 projects. There are also outliers in the data which need special attention. For example, there are volunteers with too many skills/interests, let’s say around100 skills or interests.

Chapter 2

# Related Work

The main objective of this project focuses on implementing a recommendation system based on the Vollie's dataset, which can recommend projects to volunteers. This problem has the same concepts as the user-item recommendation problem, which several papers have tackled using multiple approaches that will be discussed below.

Collaborative filtering has been a prosperous methodology using transaction information together with user and item features for recommendation. Other methodologies have been developed where the process is enhanced by mapping transactions into a bipartite user-item interaction graph and transforming the recommendation system into a link prediction problem.

In papers such as (Zhang, Li, Meng, & Zhang, 2018) and (Hasan, Chaoji, Salem, & Zaki, 2006) have concluded that the use of bipartite graph improves the performance of the models notably.

The two main methodologies are discussed below: Recommendation Systems and Link Prediction.

## 2.1 Recommendation System

This procedure aims to find a number of projects of interest to each user, on the basis that similar users will choose similar items and that that each volunteer will buy items similar to the ones that the user picked in the past. Among the recommendation systems methods, collaborative filtering is one of the most used that achieved good results.

The approach of collaborative filtering is used in (Zhang, Li, Meng, & Zhang, 2018), to compare the performance of this model with the novel approach implemented by the authors. The authors used the projection of the bipartite graph into one entity, to determine new links in the projection of the other entity based on the similarity of the nodes.

Moreover in (Chiluka, Andrade, & Pouwelse), the authors combine collaborative filtering with the link prediction advantages. Then, based on topological and node attributes they calculated a proximity score which helps predicting the relevance of an item to a user. The authors used 6 algorithms to calculate the proximity score and tested them one the data giving different results, evaluated on MAP.

## 2.2 Link prediction

According to (Zhang, Li, Meng, & Zhang, 2018), the purpose of this type of models is estimating the likeliness for future links between the two entities where there has been no interaction in the past, based on the observed links. This approach calculates the topologic similarity between nodes and complements this information with the characteristics of each node to predict the new links.

In (Zhang, Li, Meng, & Zhang, 2018), authors used the domain knowledge technique, taking external information in consideration to solve the product-user problem. Additionally, the bipartite graph was projected to a uni-partite graph to simplify the graph structure. Then, some similarity features are calculated between each customer-product pair where there was no link between them, like common neighbors and Jaccard index. Finally, they developed a ranking score using the features engineered to determine the possibility of the link between the pair of customer-product.

The main approach in (Allali, Magnien, & Latapy, 2013) is to project the bipartite graph into a uni-partite graph and predict the missing internal links in the graph, where an internal link is a pair of nodes such that adding the link into the bipartite graph does not change the projection. Usually performing the projection is a cause of loss of information, leading the authors to boost the resulting uni-partite graph with weighted edges, based on different measures like common neighbors, Jaccard coefficient, preferential attachment, etc. After that, prediction of a new internal link is determined by comparing the edge weight with a given threshold, and those weights greater than this value predicts a new internal link between those two nodes.

The previous models are intended to improve the feature engineering process, while there are other types of approaches that construct a graph kernel on user-item pairs that is represented with a bipartite graph structure together with the user and item features, such as (Kunegis, De Luca, & Albayrak, 2010), (Li & Chen, 2013) and (Kasper, Jimenez-Bonnet, & Zhang, 2012)

In the case of (Kunegis, De Luca, & Albayrak, 2010), the authors used a class of graph kernels called spectral information kernel obtained from the network adjacency matrix that represents the bipartite graph. This matrix is decomposed and transformed to get the spectral information. Then this information is plugged into Odd Pseudo Kernels models like hyperbolic sine, Odd von Neumann Pseudokernel or Rank reduction, which helped improving the MAP in different tested datasets.

For (Li & Chen, 2013), constructed the user-item graph from transaction histories and extracted features describing both nodes (user and items), then they designed a kernel on user-item pairs based on the features of the graph and the characteristics of the nodes. Finally, the kernel is plugged into a one-class SVM to separate potential links from impossible links. (Kasper, Jimenez-Bonnet, & Zhang, 2012) implemented the model to recommend businesses to users.

Chapter 3

# Data

The project team were given 64 csv files (craft tables) which formed the back end of the Vollie 1.0 system. These tables were non-relational in nature and consisted of identically named fields in different tables and multi-purpose fields within the same table designed specifically to support the Vollie front end. A schema was not provided to the team. Hence the first step was to identify the entities and fields related to this project. In the second step, the relationships between the raw files and the identified entities and fields was established.

## 3.1 Data structure and entity relations

After analysing the data received from Vollie, project has identified the entities and relationships among the between the entities. Resulting entity relationship diagram is shown in Figure 1: ER diagram of Vollie data.

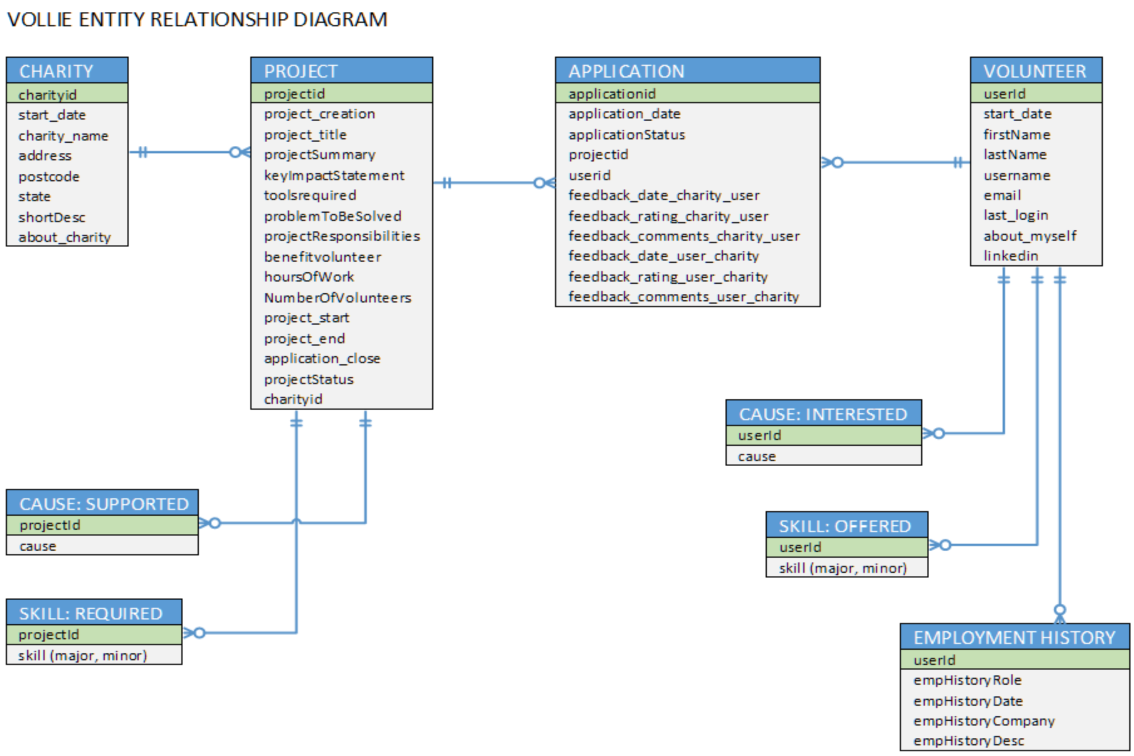


Figure 1: ER diagram of Vollie data

## 3.2 Charities

There are 452 charities registered in Vollie. The charts below show the demographic presence and how active the charities are in Vollie.

| | |
|---|---|
| Total number of charities registered in Vollie | **452** |

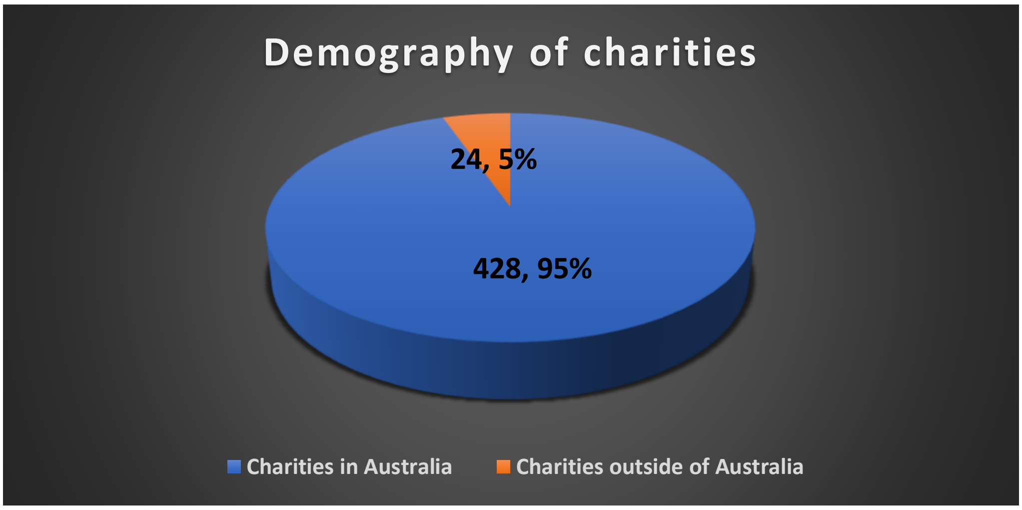


Figure 2: Demography of charities in Vollie

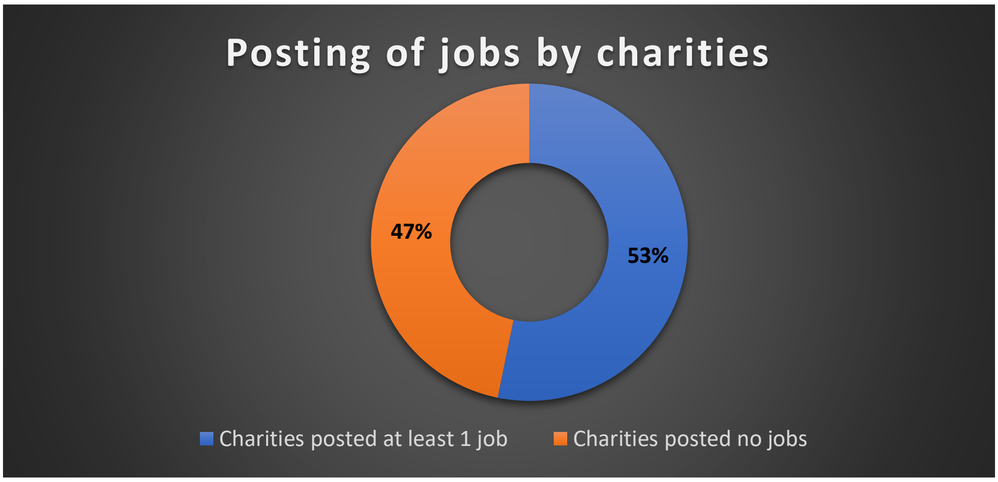


Figure 3: Posting of jobs by charities

**Observations:**

- Figure 2 shows that 95% of charities registered in Vollie are from Australia
- Figure 3 shows that 47% of charities registered in Vollie did not post even a single job.

## 3.3 Projects

According to the data received from Vollie, charities have posted 1097 jobs till date. The below table and chart describe some trends and statistics about the data related to projects in Vollie.

| | |
|---|---|
| Total number of projects | **1097** |
| Number of projects approved | **469 (42%)** |
| Number of projects pending | **515 (47%)** |
| Number of projects completed | **178 (16%)** |
| Average number of projects submitted / charity | **2.4** |
| Number of projects submitted by top 10 charities | **323 (30%)** |
| Number of projects with feedback | **42 (4%)** |
| Number of projects yet to receive application | **257 (23%)** |

Table 1: Statistics about projects

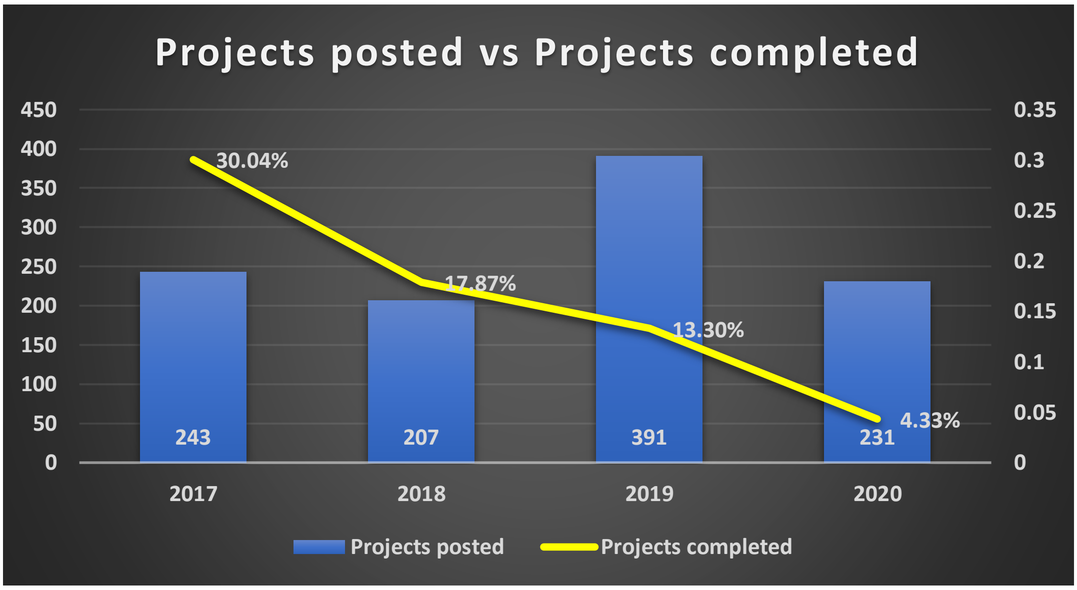


Figure 4: Project posted vs Project completed

**Observations:**

- According to the Figure 6, rate of completion of projects (no. of projects completed/number of submitted) is decreasing every year.

- As shown in Table 1, almost 47% remain in the pending stage, and this is potentially due to one of the following scenarios:
    - no application has been received for some of the projects
    - no suitable applicant or volunteer has been found for some projects
    - charity intends to put project on hold for various reasons
- As shown in Table 1, feedback related data is available for only 4% of the total submitted projects.
- As shown in Table 1, approximately 23% of projects have yet to receive a single application. This may be due to the reason that these projects never listed in the volunteers' job search result.

## 3.4 Causes

According to data received from Vollie, there are 33 fields related to causes. Charities post projects which support certain causes. Volunteers specifies the causes they are interested to support. Some interesting statistics related to caucuses are given below.

| | |
|---|---|
| Total number of fields related to causes | **33** |
| Average number of causes / projects | **2.2** |
| Median number of causes / projects | **1** |
| Average number of causes / volunteers | **2.8** |
| Median number of causes / volunteers | **1** |

Table 2: Statistics of fields related to causes

- Top 3 fields related to causes posted by projects
    1. Community Engagement (10.58%)
    2. Children (9.80%)
    3. Animal Welfare Conservatism (8.98%)

- Top 3 fields related to causes supported by volunteers
  1. Mentoring (8.03%)
  2. Animal Welfare Conservatism (7.92)
  3. Children (7.1)

**Observations:**

- 2 out of top 3 causes supported by projects matches with 2 out of top 3 causes in volunteers are willing to support.

## 3.5 Skills

According to data received from Vollie, there are 16 fields related to major skills and 85 minor skills Charities post projects which support certain causes. Charities specify skills required for projects when at the time of posting. Volunteers specify skills in their profile. Some statistics related to skills are given below.

| | |
|---|---|
| Total number of major fields related to skills | **16** |
| Total number of minor fields related to skills | **85** |
| Average number of major fields related to skills / project | **1.6** |
| Median number of major related to skills / project | **1** |
| Average number of minor related to skills / project | **2.6** |
| Median number of minor related to skills / project | **2** |

Table 3: Statistics related to skills in Vollie

- According to data, top 3 fields related to major skills required in projects
  1. Marketing & Communications (29.64%),
  2. Business (12.75%)
  3. Design (11.65%)

- According to data, top 3 fields related to major skills offered by volunteers
    1. Marketing & Communications (22.53%)
    2. Business (12.61%)
    3. Media and Arts (9.87%)

**Observations:**

- Top 2 skills in demand by charities matches with top 2 skills of volunteers in Vollie.

## 3.6 Volunteers

Volunteers are registered as There are approximately 2683 volunteers registered in Vollie. Volunteers create profiles, apply for job, works on a project if their application gets selected.

| | |
|---|---|
| Total number of volunteers | **2683** |
| Number of volunteers applied to at least one job | **960 (36%)** |
| Number of successful volunteers | **379 (14%)** |
| Number of successful volunteers having LinkedIn profile | **301 (79%)** |

Table 4: Statistics about volunteers

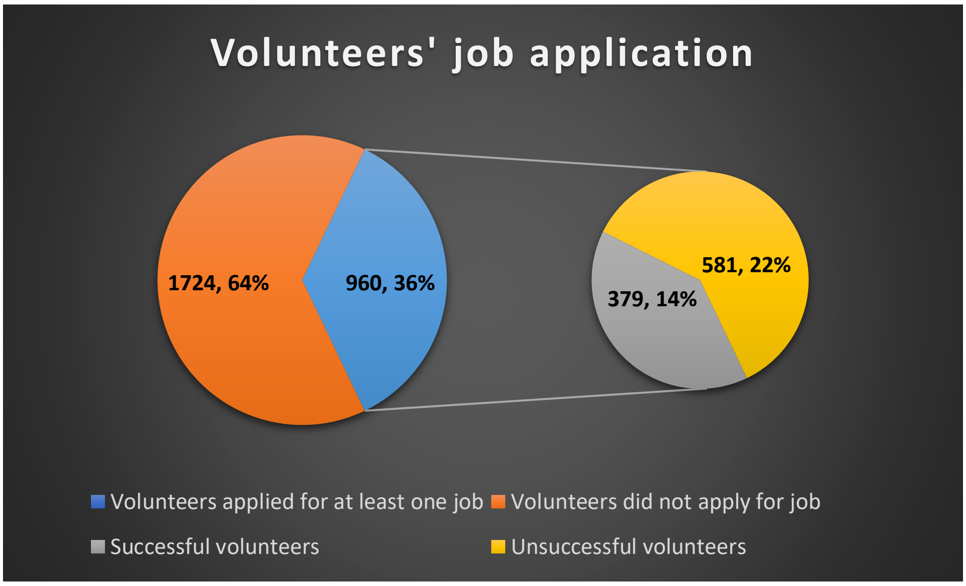

Figure 5: Volunteers' job application proportions

**Observations:**

- As shown in Table 4, maximum 79% of the volunteers who have at least 1 approved project have LinkedIn profile.
- As shown in Figure 5, only 36% of volunteers registered in applied for at least one job and 14% of the volunteers are successful to secure a job.
- According to received data, only 37% of volunteers have listed their previous employment history (previous roles, dates, companies and short descriptions)
- Charitable cause related data is not available for 39% of volunteers
- Major category skills related data is not available for 39% of volunteers

## 3.7 Applications

Volunteers submit applications to apply for jobs posted by charity. Charities screens the applications to select the volunteer to award the project.

| | |
|---|---|
| Total number applications | **1874** |

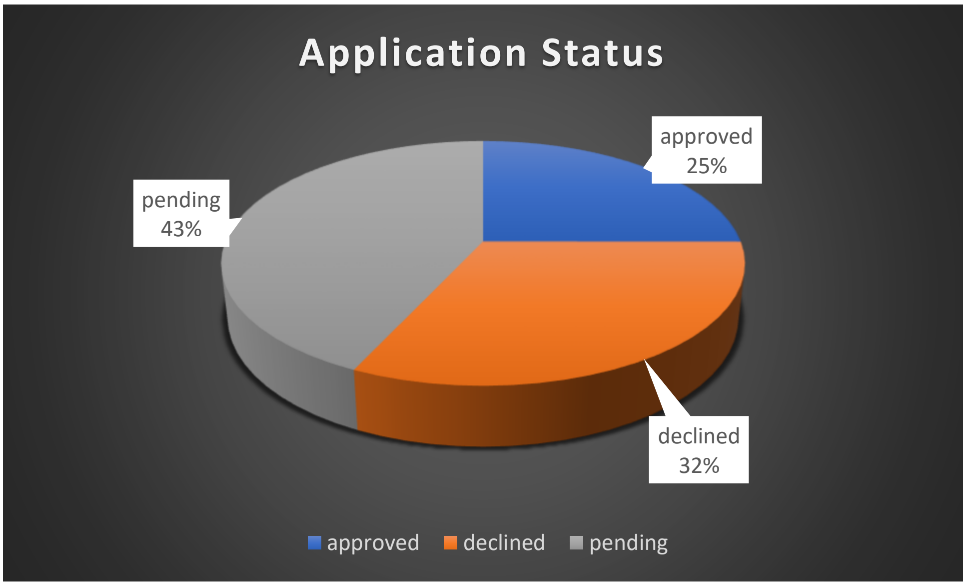


Figure 6: Application status

**Observations:**

- As shown in Figure 6, 43% of the applications are in “pending” status. That indicates that charities did not spend time and effort to process these applications.

Chapter 4

# Method

Vollie dataset is composed of charities, projects and volunteers as explained in the data section. For each charity there is information about the mission and vision; for projects there is information about the description of the projects and the skills required; and, for the volunteers there is information about its description, LinkedIn profile and skills acquired.

The framework proposed in this document is composed by four steps.

1) Creating a Bipartite graph between volunteers and projects. Also extract features describing the volunteer-projects interaction and project-project interaction as well.
2) Evaluate four main approaches that will allow the system to recommend project to volunteers: Collaborative filtering, Collaborative filtering with link prediction, Internal link prediction and graph kernels.
3) Models are assessed based on some evaluation metrics.
4) The model with the best performance is selected and ready for deployment.

As a first step, a bipartite graph is constructed to leverage the graph structure, where volunteers and projects are the nodes with edges corresponding to a project selected and carried out by a volunteer. Then, topological features are obtained from the bipartite graph using methods such as random walk, preferential attachment, or common neighbors. Moreover, features comparing the characteristics of projects-volunteers and projects-projects like Jaccard index can also be utilized.

The second step is responsible of constructing and tuning the four main approaches. A collaborative filtering will be built using only the features of the volunteers and the projects together with the transactional data. This model will be the first approach to solve the problem and will be taken as the baseline. For the collaborative filtering with link prediction, similar to what is proposed in (Chiluka, Andrade, & Pouwelse), multiple algorithms will be constructed to get indexes and theses will be engaged to get a general score, that with the help of a threshold will determine where there will be a new link or not. For the internal link approach, the bipartite graph will be projected into a uni-partite graph and the features will represent in this case the weights in the edges which will be compared against a threshold too. Finally, a graph kernel is calculated using both bipartite graph and external features. The resulting kernel will be plugged in a one-class SVM to predict new links. In this method spectral information can also be obtained from the graph kernel similar to approach proposed in (Kunegis, De Luca, & Albayrak, 2010).

The proposed models will be evaluated in the third step. Given that these approaches are modeled as a one-class classification, where there exist a link between two nodes or not, the recall, precision and F1-score will be calculated to assess the performance of each model.

Finally, the best model will be selected based on its performance.

Chapter 5

# Timeline

The project team has a 12 weeks plan to design, build, evaluate approaches proposed in Section 4. Team will finally select the best approach and will make it ready for deployment. Finally, written report will be submitted and a demonstration will be presented.

## 5.1 Timeline

In order to have a clear plan, a timeline of next semester has been shown here.

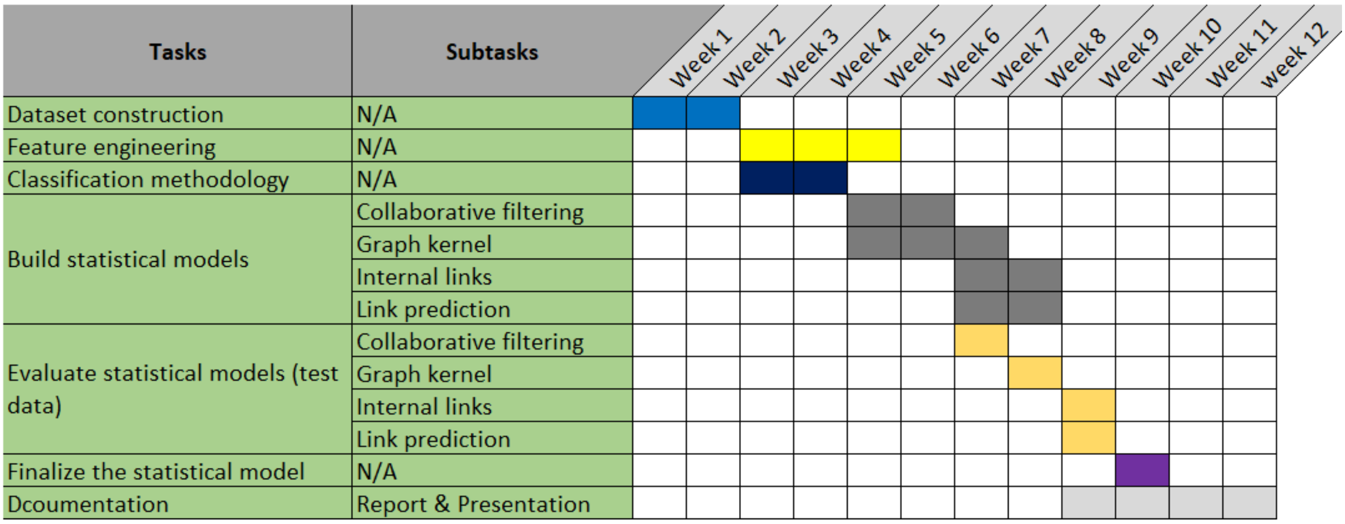


Figure 7: Project timeline including milestones

**Dataset construction:** This step involves the following activities.

- Exploring opportunities to generate/gather more data.
- Cleaning up outliers

**Feature Engineering:** This step involves the following activities.

- Building the features
- Preparing final data set as input to statistical models

**Classification methodology:** This step is required to design a model to process new project and new volunteers. For example, create the clusters of projects.

**Build statistical models:** The four statistical models (Collaborative filtering, Graph kernel, Internal links, Link prediction) will be built using the training dataset.

**Evaluate statistical models:** The models will be evaluated using test dataset against both existing volunteers, existing projects and new volunteers, new projects. Measures like precision, F-score will be used to compare the models. For evaluation of models test data set will be used.

**Finalize the statistical model:** After evaluation of four models, the best performing model will be selected and will be executed against the production data set. The performance will be noted and a final conclusion will be derived.

**Documentation:** The final project report and a presentation will be documented.

Chapter 6

# Appendix

## 6.1 Charity causes

Advocacy
Aged Care
Animal Welfare/Conservation
Awareness and Research
Cardiac Emergency Response
Children
Clinical Research
Community Development
Community Engagement
Disability Support
Disease Prevention
Domestic violence reduction
Education
Environmental Conservation
Family Support
First Aid
Health
Human Trafficking
Hunger Relief
Indigenous group support
Indigenous land management
Mental Health
Mental Illness
Mentoring
Missing Persons Support
Palliative services
Poverty Alleviation
Road Safety
Social Enterprise
Sustainable Tourism
Women's Empowerment
Youth Empowerment
Youth Services

## 6.3 Skills

- Accounting
  - Accountant
  - Bookkeeper
- Administration
  - Data Entry
  - Administration
  - General Admin
- Big Data
  - Data Analysis
  - Data Mining
- Business
  - Sales Strategy
  - Business Development Management
  - Corporate Partnerships Management
  - Leadership
  - CEO and General Management
  - Business Strategy
  - Board Advisory
  - Government Policy, Planning and Regulation
  - Business Analyst
- Design
  - Graphic Design
  - Fashion and Textile Design
- Digital
  - Website Development
  - Mobile Development
  - Analytics Support
  - App Development
  - UI Designer
  - UX Designer
- Education
  - Education Services
- Finance
  - Financial Reporting
  - Financial Strategy and Planning
  - Financial Management
- Healthcare and Medical
  - Psychology
- Human Resources
  - HR Management
  - Workplace Training and Development
  - OH&S
  - Organizational Development
  - Recruitment Strategy
  - Recruitment Delivery
- Information Technology
  - IT
  - Software Tester
- Legal
  - Legal Policy
  - Banking and Finance Law
  - Construction Law
  - Corporate and Commercial Law
  - Environmental Planning Law
  - Industrial Relations and Employment Law
  - Insurance and Superannuation Law
  - IP Law
  - Legal Assistant
- Marketing and Communications
  - Copywriting
  - Marketing Strategy
  - Brand Management and Strategy
  - Advertising
  - Social Media

- Direct Marketing
- Marketing Management
- Content Strategy
- Creative
- Search Engine Optimisation (SEO)
- Events
- Communications Strategy
- Market Research
- Campaign Management
- Sponsorship
- CRM
- Internal Communications
- Marketing Coordinator
- Digital Marketing
- Market Analysis
- Search Engine Marketing (SEM)
- Proofreading

Media and Arts

- Public Relations / PR
- Videography
- Messaging Strategy
- Media Planning
- Illustration and Animation
- Photography
- Media Strategy
- Editing and Publishing
- Journalism
- Media Buying
- Voice-Over Artist

Operations

- Operations Management Support
- Customer Service and Management

Sales

- Fundraising Management
- Sales Analysis and Reporting
- Sales Management
- Sales Account and Relationship Management

Translation

Crowdfunding

Supply Chain Management

- Procurement

# Bibliography

Cambridge Spark. (2019, November 28). Tutorial: Practical Introduction to Recommender Systems. Retrieved from https://blog.cambridgespark.com/tutorial-practical-introduction-to-recommender-systems-dbe22848392b

Deng, H. (2019, December 05). Recommender Systems in Practice. Retrieved from https://towardsdatascience.com/recommender-systems-in-practice-cef9033bb23a

Finnerty, S., Pourshahid, A., Milne, J., & Lawrence, S. (2017, September 11). 5 steps to setting up a recommender system. Retrieved from https://www.klipfolio.com/blog/recommender-system

Fritz, J. (2020, April 09). Why Engage in Virtual Volunteering? Retrieved from https://www.thebalancesmb.com/becoming-a-virtual-volunteer-4138357

Joshi, P., & Analytics Vidhya. (2020, April 21). Link Prediction: Link Prediction in Social Networks. Retrieved from https://www.analyticsvidhya.com/blog/2020/01/link-prediction-how-to-predict-your-future-connections-on-facebook/

Pandey, P. (2019, May 25). Recommendation Systems in the Real world. Retrieved from https://towardsdatascience.com/recommendation-systems-in-the-real-world-51e3948772f3

Real Python. (2019, July 22). Build a Recommendation Engine With Collaborative Filtering. Retrieved from https://realpython.com/build-recommendation-engine-collaborative-filtering/

Sharma, P. (2019, September 04). Comprehensive Guide to build Recommendation Engine from scratch. Retrieved from https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-recommendation-engine-python/

Sun, L. (2018, May 22). UX Case Study - Improving Vollie's online volunteering experience. Retrieved from https://medium.com/@danlisun/ux-case-study-improving-vollies-online-volunteering-experience-5d4eaa06c20f

Vollie Pty. Ltd. web site https://www.vollie.com.au/blog/2017/01/19/non-profits-dont-need-money-they-need-volunteers/

Customer-Product Bipartite Graph for Product Recommendation. International Journal of Information Technology and Decision Making.

Hasan, M., Chaoji, V., Salem, S., & Zaki, M. (2006). Link Prediction Using Supervised Learning.

Chiluka, N., Andrade, N., & Pouwelse, J. (n.d.). A Link Prediction Approach to Recommendations in Large-Scale User-Generated Content Systems. In S. B. Heidelberg, Advances in Information Retrieval (pp. 189-200).

Allali, O., Magnien, C., & Latapy, M. (2013). Internal link prediction: A new approach for predicting links in bipartite graphs. Intelligent Data Analysis.

Kunegis, J., De Luca, E. W., & Albayrak, S. (2010, June). The Link Prediction Problem in Bipartite Networks.

Li, X., & Chen, H. (2013). Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach. Decision Support Systems.

Kasper, B., Jimenez-Bonnet, S., & Zhang, W. (2012, December). Personal Recommendation as Link Prediction using a Bipartite Graph Kernel.