|  |  |
| --- | --- |
| Meeting Minutes 12 | |
| Date | 13 January 2020 |
| Start Time | 10am |
| End Time | 11am |

|  |  |  |
| --- | --- | --- |
|  | Agenda | Follow Up Action |
| 1 | Consultation with supervisor, Prof Akshat, to run through objectives and plan of handling project.   * Clarify expectations * Gather feedback on data cleaning and model building | Next meeting with Prof Akshat: 22 January 10:30am – 11:30am |

Notes from meeting:

**Tanny**

* Unsupervised learning is a very good technique used on many deep learning can visualise higher dimension data. Can use unsupervised learning tools to explore.
* TSNE - popular (no supervised data, automatically clusters data, distinct clusters based on data) can be used for survivability or cost prediction based on the clusters
* Text mining to generate feature automatically from the data, discard stop words etc. Called TOPIC modelling - tool called Mallet can convert text into features (if have sufficient time to do it as it will be useful and can be very meaningful, gives details)
* Deep learning only works because people have access to high quality data
* Interesting to see basic description like K-means clustering etc. Like basic Descriptive data models
* Show a range, not necessary flat numbers for cost calculation output, 95% CI, means, median, SD
* Generate all useful feature, put into deep learning tools and let them identify what is the best tools for value prediction, cost prediction as we have features for the historical cost.
* Out of 100k how much is spent on which area (chemo, surgery etc.) on average.
* Key point: **Accuracy** (mean squared error) of the results should be the goal and score, focus on accuracy, others like visualisation is secondary.
* Spend more time on data processing, feature extraction.
* 1 midterm presentation (wk 8), 1 final presentation (wk 13)
* Visualise what should be met by midterm (substantial outputs) and final presentation (invite end users, meet their expectations)
* What to show on midterm/ final term: Problem is important, how data pre-processing is done, how workload is split
* NSCC can use by VPN government’s supercomputing for Uni students have very strict PDPA with whole IT teams. Hard to do deeplearning for laptop unless gaming laptop.
* Advise: Spend more time on project gantt chart/workload split should be secondary project manager should focus on project’s results just update once in awhile can be done anytime
* Wiki not necessary for now.

**Jia Yun**

1. Do you want to drop free text or: Topical analysis that can generate features automatically for you. Stop words: enter, etc. left with some topics and keywords to describe. Topic modelling. Software: Mallet. Might not want to drop the free text if there is some unsupervised free text you can get. - free text can be useful as code cannot capture the full thing. Mallet can limit how many topics you want also.

2. How to add back: e.g ID and visit. The unique identifier.

3. Focus on getting the main thing. If got additional time can use the mallet. Would be good to explore

4. Interesting to see: basic k-means clustering. At least do more insightful charts (rather than just the number patients, basic excel charts, etc)

5. Typically people don’t like to see flat number. They like to see a range for the price. Not the exact price. E.g. $10-15k. Try to calculate variance and confidence interval. Average, median and standard deviation

6. Patient data to cluster data — two levels of approximation. However, for value and cost prediction can also let supervised deep learning decide which are the most relevant features.

7. Is it good to give unfiltered information to patients? Let them decide what to give to patients. - would patients care?

8. Most interesting thing to make our team stand out: accuracy and validation of our model. Getting data that is rich enough and have good features. Most important!!

9. When presenting results to doctors, show the validation. 9-fold validation. Randomly select 8 clusters.

10. Take enough time to do data processing and feature extraction.

11. Accuracy: in x years x amount of dollars. Or can use range. Can use mean squared error, etc.

12. one midterm presentation (week 8)

13. Final presentation (week 13)

14. Project managementt: in this kind of project cannot always meet deadlines exactly. Should visualize what you want to do by midterms and finals. If what you specified is met at those 2 major milestones, then its ok.

15. Those in acceptance grading excel can put in presentation. Make sure have equal workload.

16. His focus is on accuracy, hard number.

17. IS483 don’t need wiki page??? Not a requirement currently.

18. Meeting w prof: every Wednesday: 10:30-11:30am

19. National supercomputing centre: get account and use vpn. Can check with nccs. The nsc supposed to have some strict regulating. Increase ram of laptop. Can be difficult to get an account in nsc, might need to ask Prof Yeow Leong.