Feature Engineering for a Building a Machine Learning Model for Recommendation System for Generic Competency Development in Higher Education

Presented at the CPCECPR Conference 2022

January 13, 2022

Authors: Adam Wong, Joseph So, Ada PL Chan, Kia Tsang

This study was supported by the Faculty Development Scheme (No. UGC/FDS24/E09/20) of the University Grant Committee of Hong Kong

What is Feature Engineering?

- Raw data collected from various sources may not be in the suitable form that is compatible with the specific machine learning algorithm for a certain purpose.
- For machine learning models to produce useful knowledge to inform educators, the algorithms must be provided with relevant data about the students.
- Feature engineering is applied to transform raw input data so that they can be used by machine learning algorithm correctly.

Background

- A education data mining project to find out the associations between student academic performance and
  - participation in various activities
  - demographics
  - academic performance prior to joining the university
- The dataset consists of many worksheets in an Excel workbook.

B	C	D	E	F	G	
Latest Program	Final Registration S	Source of Infor	Category Code	Sub-Category (	Activity Code	Activity Title 1
8C108-HR	Graduated	CSAO	DP	PD	PD19105A-191-01	Mindfulness –
8C108-HR	Graduated	CC	OTH	SU		HKCCSU Rotar
8C121-SVM	Graduated	CSAO	A/SH	AWD		PolyU HKCC Ac
8C108-HR	Study Ended, Not Reg	CSAO	A/SH	AWD		PolyU HKCC Ac
8C111-PSY	Graduated	CC	OTH	TK	TK18110P-181-01	Research Sem
8C111-PSY	Graduated	CC	A/SH	AWD		CPCE Dean's Li
8C111-PSY	Graduated	CC	DP	PD	PD18123P-182-01	Introduction t
8C111-PSY	Graduated	CC	OTH	TK	TK13215S-182-01	Forensic Psych
8C111-PSY	Graduated	CSAO	A/SH	AWD		HKCC Director
8C111-PSY	Graduated	CSAO	A/SH	SCH		PolyU HKCC Oı
8C111-PSY	Graduated	CC	OTH	TK	TK19111P-191-01	Research Sem
8C111-PSY	Graduated	CSAO	DP	CS	CS16133A-191-01	Psychological
8C111-PSY	Graduated	CC	OTH	TK	TK19106P-191-01	Professional C
8C111-PSY	Graduated	CC	A/SH	AWD		CPCE Dean's Li
8C111-PSY	Graduated	CC	A/SH	AWD		CPCE Dean's Li
8C111-PSY	Graduated	CSAO	DP	FS	FS18101A-181-02	Non-JUPAS Pe
8C111-PSY	Graduated	CSAO	DP	PD	PD17101A-181-01	Self-learning L
8C111-PSY	Graduated	CC	DP	EE	EE18103P-181-01	Essay Writing
8C111-PSY	Graduated	CC	DP	PD		Visual Journey
8C123	Graduated	CC	DP	PD		Psychodrama i
8C123	Graduated	CC	A/SH	AWD		CPCE Dean's Li
8C123	Graduated	CC	DP	PD		Mindfulness a
8C123	Graduated	CSAO	DP	CS	CS16133A-181-01	Psychological

WIE_requirement | cumulative_GPA_by _sem | grade for all subjects

B	C	D	E	F	G	H
Latest Program	Final Registrati	Source of Infor	Category Code	Sub-Category (	Activity Code	Activity Title 1
8C108-ACC	Officially	CC	DP	PD		HKCC Orientation 2019 - To
8C108-ACC	Officially	CC	DP	PD		HKCC Orientation 2019 - To
8C108-ACC	Officially	CC	DP	PD		HKCC Orientation 2019 - To
8C108-ACC	Officially	CC	DP	PD		HKCC Orientation 2019 - To
8C108-ACC	Officially	CC	DP	PD	PD19101-191-01	Rotaract Leadership Traini
8C108-ACC	Officially	CC	DP	PD		HKCC Orientation 2019 - To
8C108-ACC	Officially	CC	DP	PD		HKCC Orientation 2019 - To
8C112-IT	Graduated	CC	DP	PD		Introduction to Python Pro
8C112-IT	Graduated	CSAO	A/SH	SCH		PolyU HKCC Outstanding S
8C112-IT	Graduated	CSAO	A/SH	AWD		HKCC Director's List 2020
8C112-IT	Graduated	CC	DP	PD		Introduction to R Program
8C112-IT	Graduated	CC	A/SH	AWD		CPCE Dean's List (Semeste
8C112-IT	Graduated	CC	A/SH	AWD		CPCE Dean's List (Semeste
8C112-IT	Graduated	CC	A/SH	AWD		CPCE Dean's List (Semeste
8C108-HR	Graduated	CC	DP	CD		Let Our Team Shine - Colla
8C108-NS	Graduated	CSAO	DP	CS	CS18107A-191-01	Elementary Korean and Ko
8C108-NS	Graduated	CSAO	A/SH	AWD		PolyU HKCC Academic Pro
8C108-HM	Study Ended,	CSAO	DP	EE	EE19103A-191-01	IELTS Series: Intensive Wo
8C108-HM	Study Ended,	CC	DP	PD		HKCC Orientation 2019 - To
8C108-HM	Officially	CC	DP	ST	ST19101P-191-01	Study Tour to Walailak Uni
8C108-HM	Officially	CSAO	DP	CS	CS18107A-191-01	Elementary Korean and Ko
8C108-HM	Officially	CC	DP	PD		HKCC Orientation 2019 - To

dent List | HKDSE Result | Subjects | WIE | WIE_requirement | cumulative_GPA_by _sem | grade

Background

- Enormous amount of data
- Total of 27763 records, not counting records in other spreadsheets

B	C	D
Latest Program	Final Registration S	Source of
8C108-HR	Graduated	CSAO
8C108-HR	Graduated	CC
8C121-SVM	Graduated	CSAO
8C108-HR	Study Ended, Not Reg	CSAO
8C111-PSY	Graduated	CC
8C111-PSY	Graduated	CC
8C111-PSY	Graduated	CC
8C111-PSY	Graduated	CC
8C111-PSY	Graduated	CSAO
8C111-PSY	Graduated	CSAO
8C111-PSY	Graduated	CC
8C111-PSY	Graduated	CSAO
8C111-PSY	Graduated	CC
8C111-PSY	Graduated	CC

WIE_requirement ...

Count: 13907 Sum: 2.52982E+11

B	C	
Latest Program	Final Registrati	Source c
8C108-ACC	Officially	CC
8C108-ACC	Officially	CC
8C108-ACC	Officially	CC
8C108-ACC	Officially	CC
8C108-ACC	Officially	CC
8C108-ACC	Officially	CC
8C108-ACC	Officially	CC
8C112-IT	Graduated	CC
8C112-IT	Graduated	CSAO
8C112-IT	Graduated	CSAO
8C112-IT	Graduated	CC
8C112-IT	Graduated	CC
8C112-IT	Graduated	CC
8C112-IT	Graduated	CC

dent List HKDSE Result ...

e: 19202445.97 Count: 13858 Sum: 2.66088E+11

Background

- Not all the features are important or needed to produce an acceptable level of confidence in performing prediction, clustering, regression, etc.
- The more features added to the model, the higher increase in time and complexity of the machine learning algorithms.
  - Ramasubramanian K., Singh A. (2017)
- The act of obtaining features from raw data and converting them into forms appropriate for machine learning models is referred to as feature engineering.
  - Zheng, A., & Casari, A. (2018)
- In other words, feature engineering is an optimization process for raw data to make them fit in the machine learning algorithms.

Challenges

- Total duration of participation is recorded in hours, days, weeks, months, and years.
- No record of total number of activities a student joined.
- Orientation-related activities make up a big proportion of one sub-category activities, i.e. “PD” activities.
- Number ranges can only be stored as text which is hard to be used by Python.
- Recording student’s GPA by year is confusing, making data preprocessing and analysis clumsy.

Challenges (Orientation-related activities)

- There are 5055 records of the sub-category “PD”.

Latest Program	Final Registration S	Source of Infor	Category Code	Sub-Category (
8C108-HR	Graduated	CSAO	DP	PD
8C111-PSY	Graduated	CC	DP	PD
8C111-PSY	Graduated	CSAO	DP	PD
8C111-PSY	Graduated	CC	DP	PD
8C123	Graduated	CC	DP	PD
8C123	Graduated	CC	DP	PD
8C123	Graduated	CC	DP	PD
8C123	Graduated	CC	DP	PD
8C123	Graduated	CSAO	DP	PD
8C123	Graduated	CC	DP	PD
8C123	Graduated	CC	DP	PD
8C108-HR	Graduated	CSAO	DP	PD
8C123	Graduated	CC	DP	PD
8C123	Graduated	CC	DP	PD
8C123	Graduated	CC	DP	PD
8C123	Graduated	CC	DP	PD
8C123	Graduated	CC	DP	PD
8C123	Graduated	CC	DP	PD
8C123	Graduated	CC	DP	PD
8C123	Graduated	CC	DP	PD
8C113-EI	Graduated	CC	DP	PD

WIE | WIE_requirement | cumulative_GPA_by _sem | g ...

Ready 5055 of 13906 records found Count: 5056 100%

Challenges (Orientation-related activities)

- However, over half of them are orientation-related activities, i.e. 2954/5055 records.
- It makes up a big part of data as students may think orientation-related activities are mandatory.
- Making the “PD” sub-category an outstanding target variable.

	D	E	F	G	
1	Source of Infor	Category Code	Sub-Category (	Activity Code	Activity Title 1
57	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
63	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
69	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
70	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
71	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
73	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
74	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
75	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
76	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
97	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
98	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
137	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
138	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
139	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
158	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
159	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
186	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
188	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
193	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
195	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses
196	CC	DP	PD		HKCC Orientation 2018 - Topic Sharing Ses

WIE | WIE_requirement | cumulative_GPA_by _sem | g ...

Ready 2954 of 13906 records found Count: 2955 100%

Challenges (Recording GPA)

- Cumulative GPA was originally recorded by year and then by semester.
- It is not ideal as it creates confusion in the data preprocessing and analysis processes.
- A more proper way to record cumulative GPA is to record from a student’s perspective.
- Normally, a student will study for 2 years in either HKCC or SPEED. Therefore, the student will study for a total of 4 semesters, without considering the year.

Solutions

- Unify all duration units to better represents the total duration of participation of one activity.
- Calculate the total number of activities a student joined, i.e. one-off and non-one-off activities.
- Separate the total number of orientation-related activities joined.
- Convert categorical data to be represented as integers for Python to understand.
- Record GPA from students' perspective, meaning recording GPA by semester 1 to 4 as a student normally studies for 4 semesters in total.

Results

- Duration units are unified as hours.
- The total number of activities a student joined is calculated at the preprocessing phase.
- The total number of orientation-related activities joined was calculated separately.
- Integers were used to represents categorical data so Python can completely understand.
- GPA is recorded by semester rather than by year and semester, making it easier to process future data.

Results

- The dataset contains the following features after preprocessing:
  - Programme
  - Gender
  - Highest Qualification
  - P Score
  - ENG
  - CHI
  - MATH
  - MATH_EXT_1_STA
  - MATH_EXT_2_ALGE
  - LIB_STUDY
  - CumlGPA Semester 1
  - CumlGPA Semester 2
  - CumlGPA Semester 3
  - CumlGPA Semester 4
  - And many more

Results

- There are many zeros in the preprocessed dataset.
- It is normal as no students took part in that category of activity.
- It is also to show there are no records for those features.
- Truly represents the real-world situation.

	AH	AI	AJ	AK	AL	AM
1	('Hrs-Cat', 'OTH-TK')	('Hrs-Cat', 'OTH-VS')	('Cat Total', 'ASH')	('Cat Total', 'CTRB')	('Cat Total', 'DP')	('Cat Total',
5133	0	0	0	0	1	
5134	0	0	0	0	0	
5135	0	0	1	0	3	
5136	0	0	0	0	1	
5137	0	0	0	0	0	
5138	0	0	0	0	0	
5139	0	0	0	0	2	
5140	0	0	0	0	1	
5141	0	0	0	0	0	
5142	2.5	0	0	0	16	
5143	0	0	0	0	2	
5144	0	0	0	0	3	
5145	0	0	1	0	0	
5146	0	0	0	2	0	
5147	0	0	0	0	2	
5148	0	0	0	0	0	
5149	0	0	0	0	3	
5150	0	0	0	0	0	
5151	0	0	2	0	9	
5152	0	0	0	0	3	
5153	0	0	1	0	3	
5154	1	0	1	0	10	

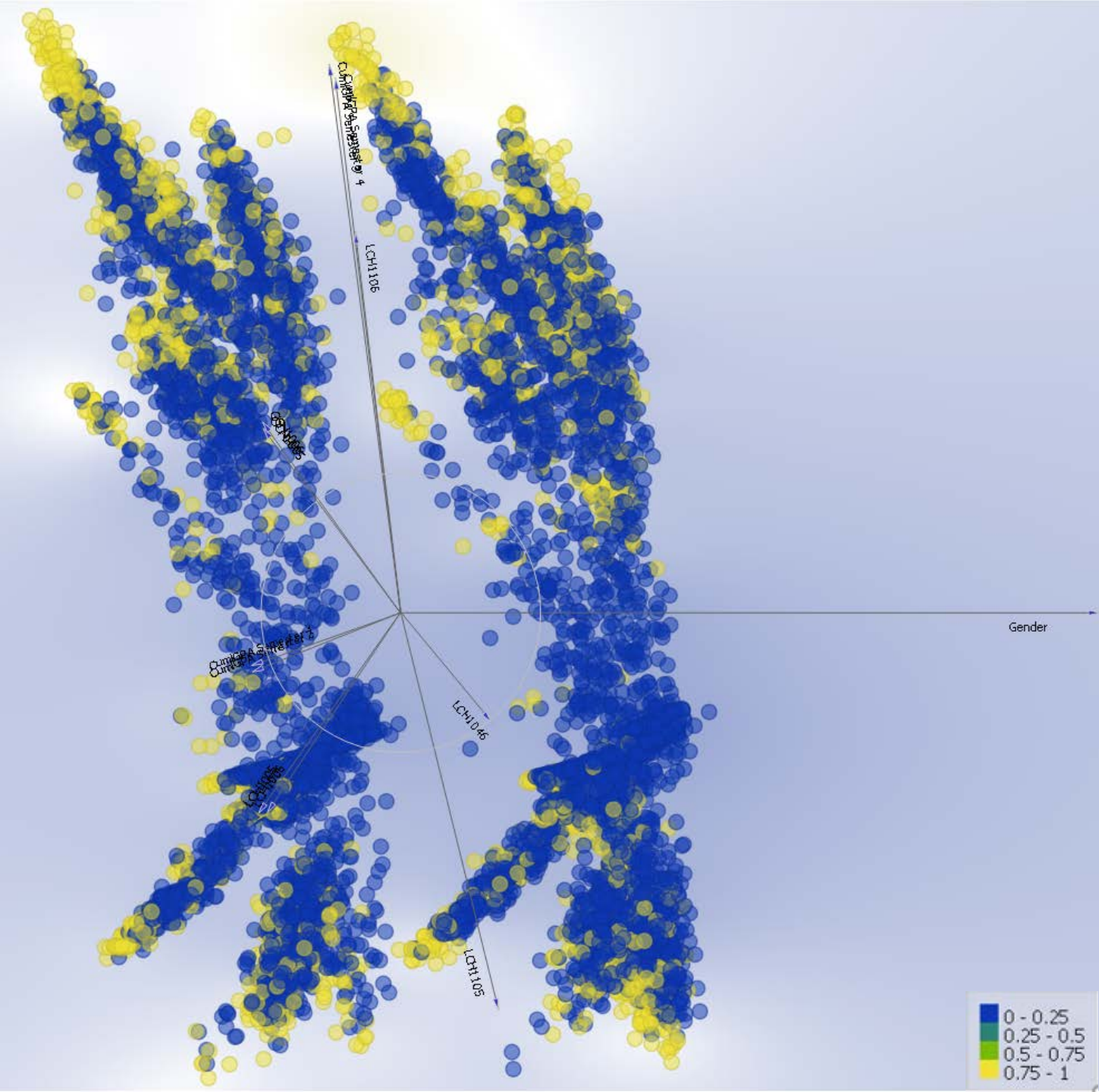
Pre-processed Data

Discussion

- After feature engineering, the data is ready to be processed by machine learning algorithms.
- The results will be applied to machine learning models.
- By implementing the solutions, we can have better understanding of the data.
- Sometimes we can even get more insights of the data.
  - Students join more one-off activities than non-one-off activities (Long-term activities).
  - Almost half of the students took part in orientation-related activities.

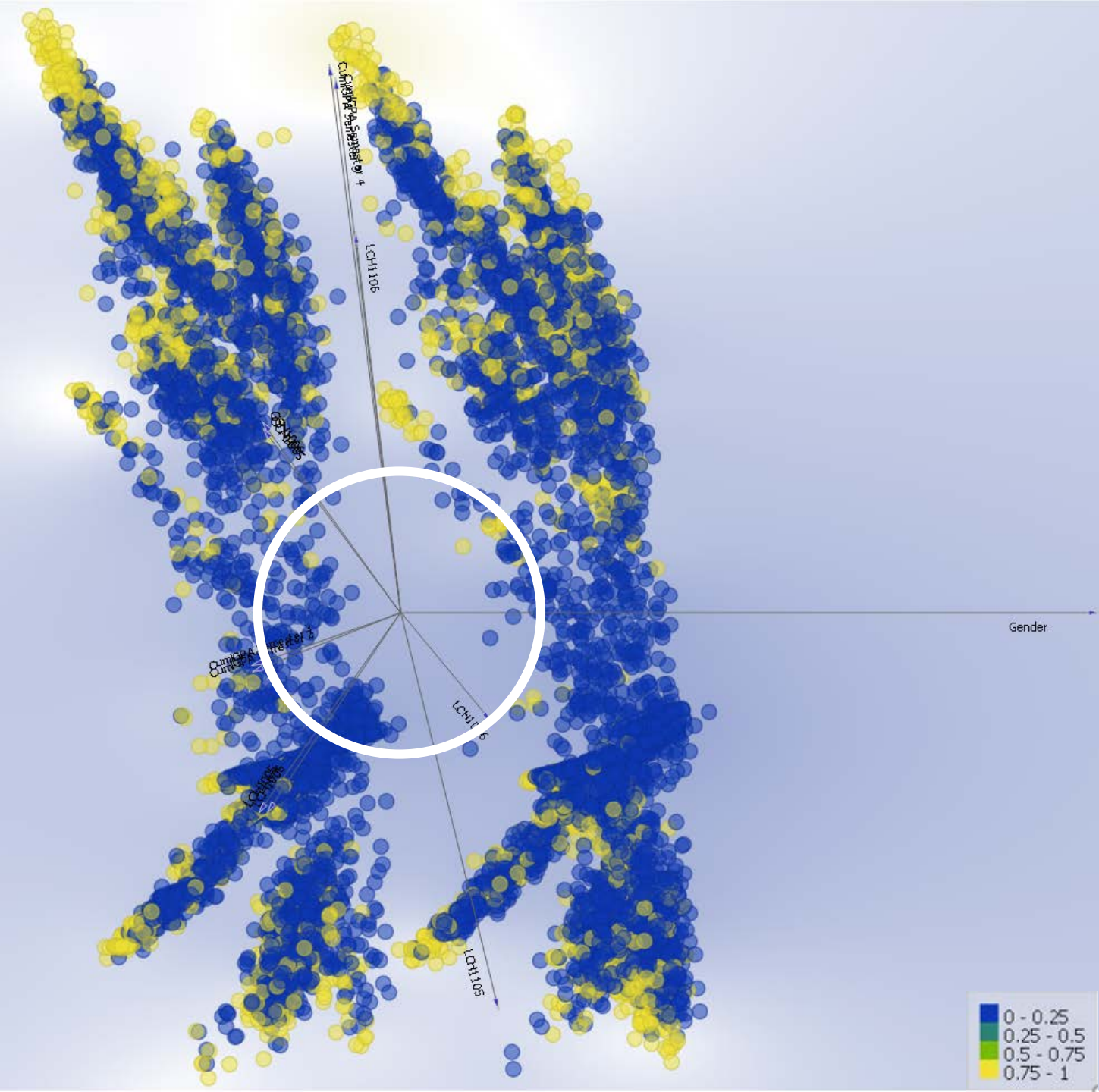
Progress

- This is a graph about principal component in AWD, meaning getting an award.
- The longer the arrow, the higher the chance of it being a principal component.
- As language subjects are the main target of the research at the current stage, the focus will be on language subjects.
- Since cumulative GPA and Gender are not language subjects, they are ignored for now. However, they are still promising principal components.



Progress

- Aside from cumulative GPA, LCH1106, LCH1005 and LCH1006 are possible principal components, which are language subjects.
- The white circle in the graph hides some features that are not as significant as others.



Next Step

- The research will be focusing on communication skills first.
- Duplicated subjects are found after updating the school curriculum. As such, subject mapping need to be done to remove duplicated old subjects.
- Activities will be recorded by sub-theme with total hours to better reflect the improvements that students can achieve.
- Self-reporting rating will be added to the dataset to record student’s self-perception.

References

- Ramasubramanian K., Singh A. (2017) Feature Engineering. In: *Machine Learning Using R*. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-2334-5_5
- Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.".

The End