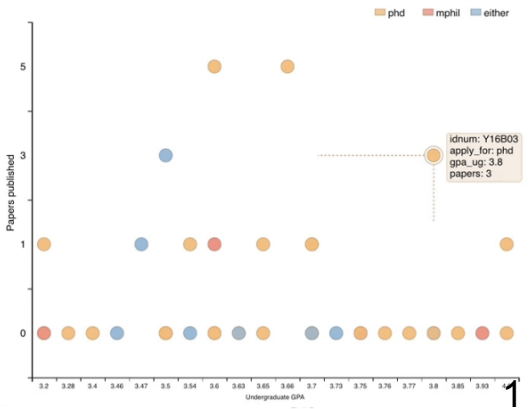
April 24th 2017

Mining HKUCS Graduate Student Data

Extraction, Analysis and Prediction

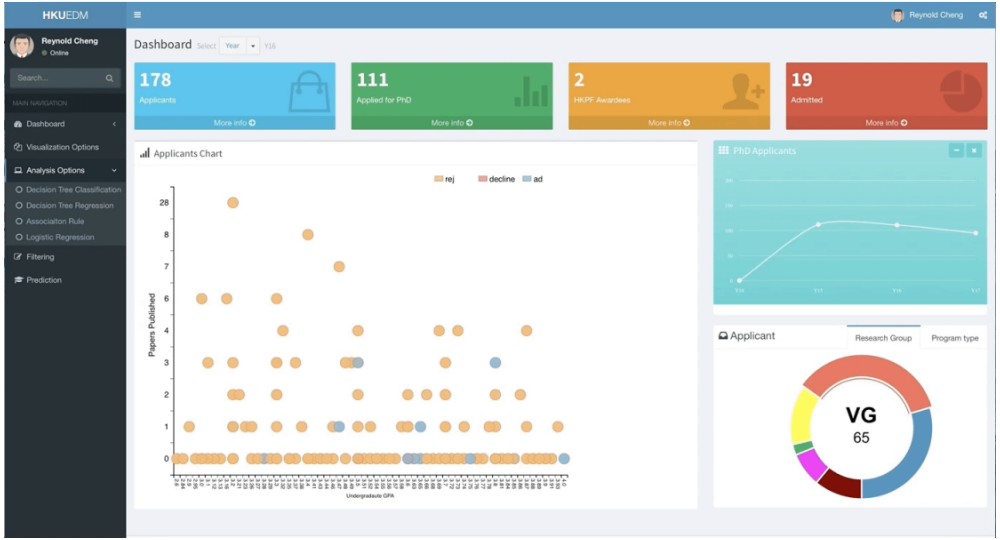
GROUP MEMBERS: WU YOU (Johnson)
XU FANGYUAN (Carrie)
SUPERVISOR: DR. REYNOLD CHENG



OUTLINE

- Web Tool Demo Video
- Methodology
- Functionalities and Implementation
- Experiments and Results
- Q & A

DEMO VIDEO



HKUEDM

METHODOLOGY

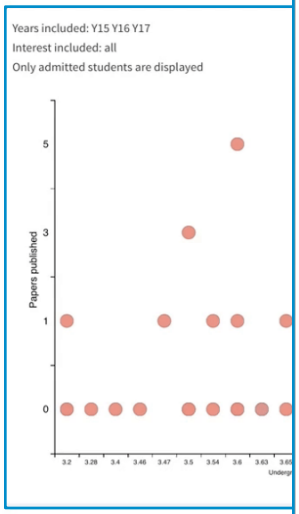
Web Framework

Database

- Data Visualization: d3.js
  - Parameter Customization
  - Interactive Functions

- Data Mining: sckit-learn
  - Model building
  - Model selection
  - Prediction

VISU



BUBBL

APPLICANT DETAIL

Applicant of 2017 Admitted HKPF Awardee

Program: PhD
Supervisor
Group:

EDUCATION

Undergraduate: Zhejiang University

Major: CS

QS Ranking: 75

81% among all applicants

GPA: 3.7700 / 4

79% among all applicant

Papers published: 0

36% among all

RESEARCH INTEREST

Interest 1: ALG

Interest 2: PL

Interest 3: N/A

ENGLISH TESTS

TOEFL :

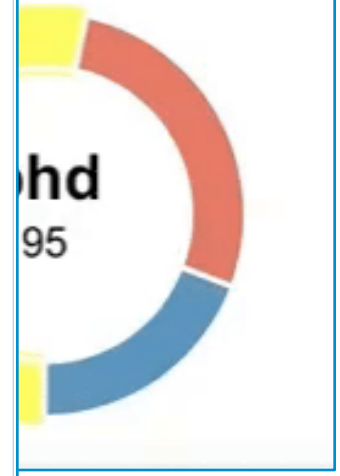
TEACHER'S COMMENTS

@cs.hku.hk

- Good English, can express herself clearly- Prefer algorithms at this point: have taken various classes in algorithms- Project in augmentation theory: can describe the ideas well- High GPA from Zhejiang U, very good chance of HKPF, - Good

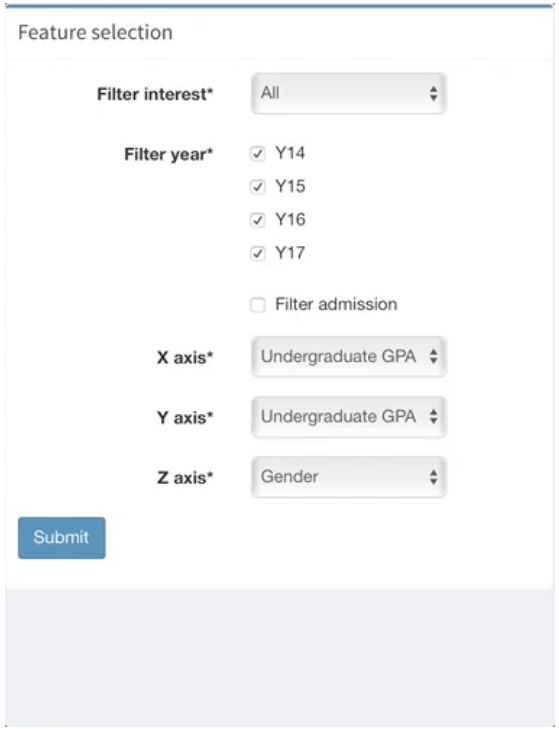
AWARDS

N/A



UT CHART

VISUALIZATION *Customization*

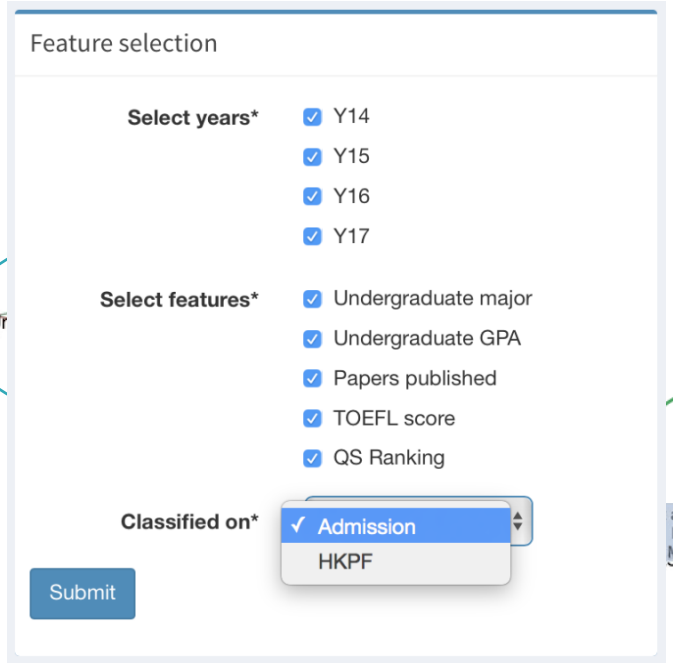


SELECT Numerical Features:
GPA, QS Ranking, Papers published, English Test Score

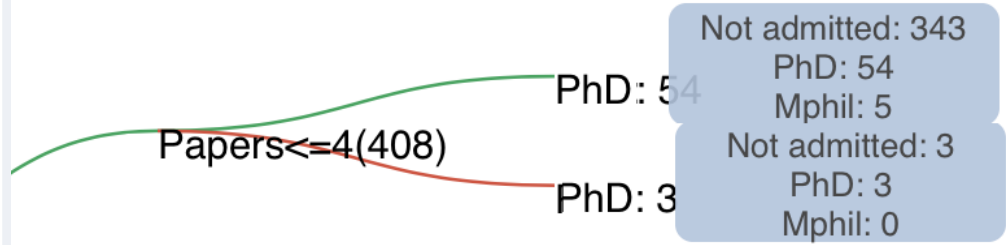
SELECT Categorical Features:
Year, Gender, Admitted Program Type, HKPF, Undergraduate Major, Research Interest, Admitted Round

FILTER Applicants:
Year, Research Interest, Admission Result

DECISION TREE *Overview*



SPLIT the dataset into different classes:
Not Admitted, PhD, MPhil

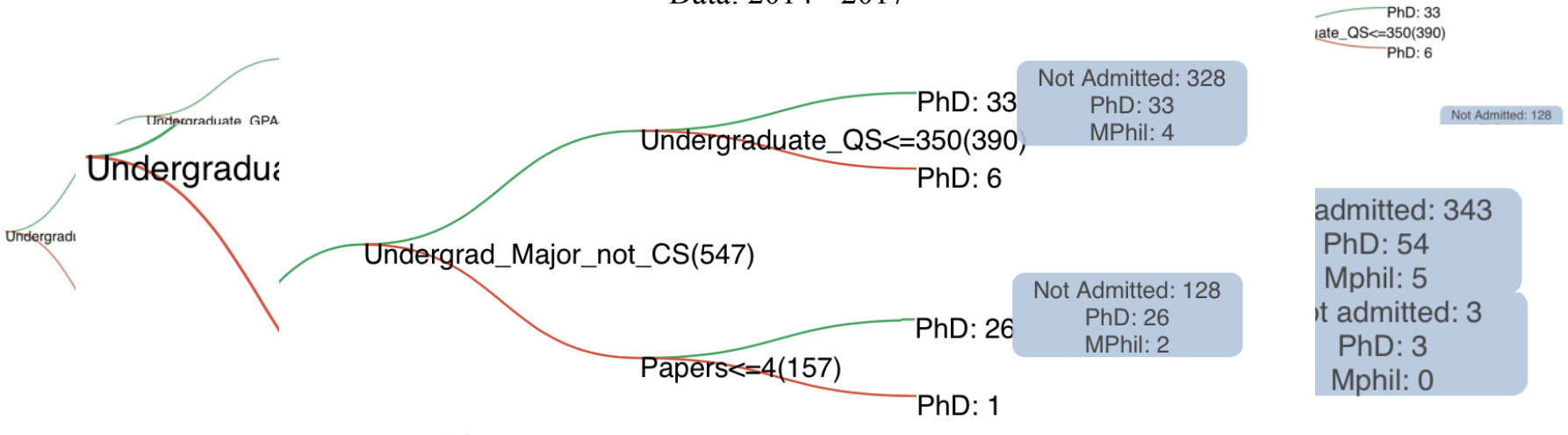


admitted: 0
PhD: 1
Mphil: 2

CUSTOMIZE Features and classes included
Admission Result
HKPF Recipient

DECISION TREE *Feature Comparison*

Data: 2014 - 2017



Undergraduate GPA > 0.86
Papers <=4

QS Ranking <=350
Undergraduate Non-CS Major

QS Ranking <=400
Papers <=4

Association Rule *Overview*

Frequent item set mining and association rule learning over transactional database.

Transactional database

- Transaction (applicant)
- Item (binary attributes)
  - e.g. GPA>0.8

Association Rule

- X -> Y
  - X is an item set
  - Y is an item
    - **support** $\text{supp}(X)/N$
    - **confidence** $\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X)$

Association Rule *Implementation*

Apriori(T, ϵ)
$L_1 \leftarrow \{\text{large } 1-\text{itemsets}\}$
$k \leftarrow 2$
while $L_{k-1} \neq \emptyset$
$C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\} - \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \not\subseteq L_{k-1}\}$
for transactions $t \in T$
$C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}$
for candidates $c \in C_t$
$count[c] \leftarrow count[c] + 1$
$L_k \leftarrow \{c \mid c \in C_k \wedge count[c] \geq \epsilon\}$
$k \leftarrow k + 1$
return $\bigcup_k L_k$

Apriori

- Bottom up approach
  - extend frequent subsets
- Breadth-first search
- Hash tree structure
  - count item set efficiently

Association Rule ***Result***

Association Rule Result

itemset	support	item base	item add	confidence	explanation
{papers < 4, qs_ug >= 112, reject}	0.64	['qs_ug >= 112', 'reject']	['papers < 4']	0.91	Itemset with highest support and its highest confidence rule
{admit, norm_gpa_ug >= 0.86, papers < 4}	0.1	['norm_gpa_ug >= 0.86', 'papers < 4']	['admit']	0.15	Highest confidence rule and its support value

Rule:

{item set}

- support

{item base} -> {item add}

- confidence

Example rule:

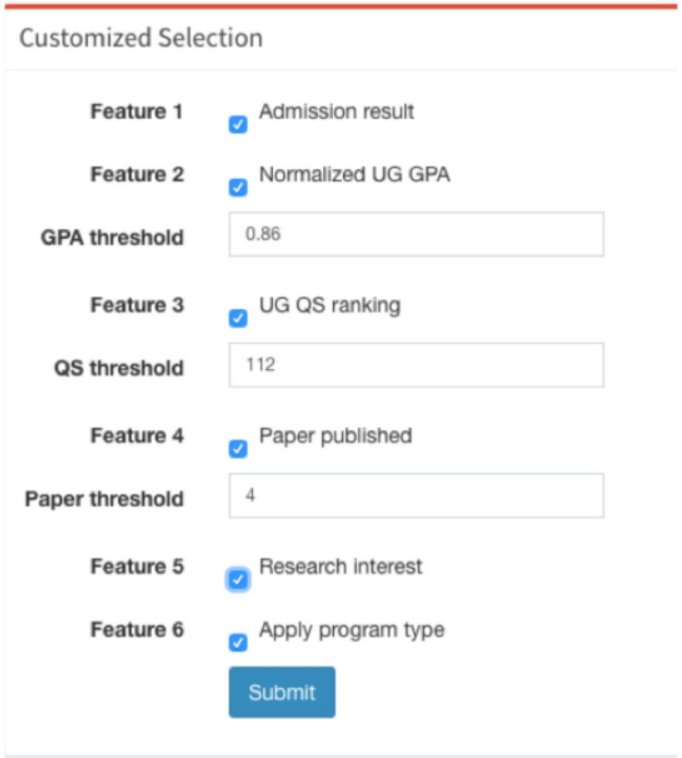
{admit, UG gpa >= 0.86, papers < 4}

- support = 0.1

{UG gpa >= 0.86, papers < 4} -> {admit}

- confidence = 0.15

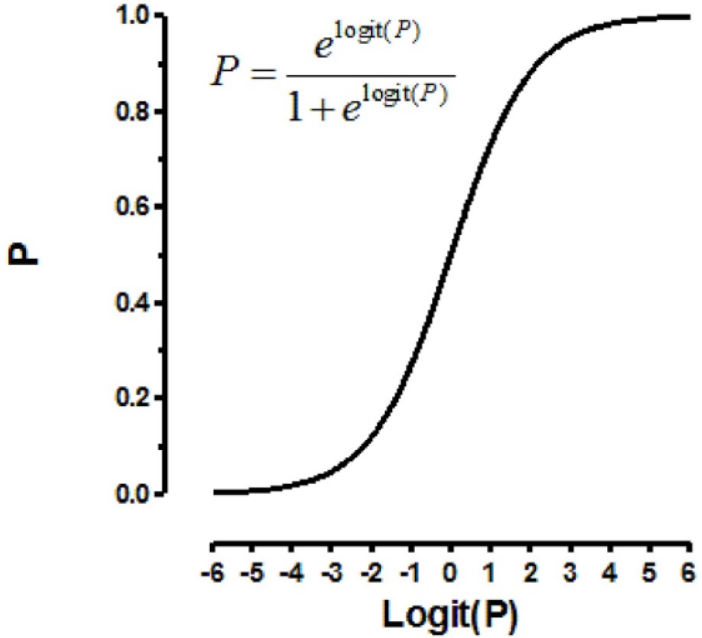
Association Rule ***Customization***



CUSTOMIZE

- Attributes to be included
- Threshold value

Logistic Regression *Overview*



$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

- **Categorical dependent variable**
- **Probability (P)**
- **Coefficient (β_k)**

Logistic Regression *Implementation*

Regression Model Result

Decision Function: reject = 0.332apply_phd + 1.5apply_mph + 0.001qs_ug + -0.31major_cs + 0.722attend_pg + 0.037papers + -0.091norm_gpa_ug + 0.524

class	apply_phd	apply_mph	qs_ug	major_cs	attend_pg	papers	norm_gpa_ug	intercept
reject	0.332	1.5	0.001	-0.31	0.722	0.037	-0.091	0.524

sklearn.linear

- Binary Classification
  - Logistic function
- Multi-class Classification
  - One-vs-rest

Logistic Regression ***Result***

ID	Predicted Result	Rank	Admit Probability	apply_phd	apply_mph	qs_ug	major_cs	attend_pg	papers	norm_gpa_ug
Y16A01	reject	68	0.163	-	mphil	15	CS	-	0	0.95
Y16A02	reject	77	0.151	phd	-	225	others	Tsinghua University	1	0.8
Y16A03	reject	112	0.113	phd	-	650	others	Tsinghua University	1	0.9125
Y16A04	reject	34	0.311	phd	-	15	others	-	0	0.82
Y16A05	admit	1	0.383	phd	-	15	CS	-	0	0.9125
Y16A06	reject	69	0.162	phd	-	175	others	Tsinghua University	0	0.823
Y16B01	reject	116	0.111	phd	-	650	others	Peking University	1	0.725
Y16B02	reject	33	0.312	phd	-	16	others	-	0	0.8575
Y16B03	reject	90	0.139	phd	-	650	CS	Peking University	3	0.95
Y16B04	reject	136	0.105	phd	-	650	others	Peking University	3	0.825

- Class probability
- Predicted Result
- Independent variable value

Logistic Regression ***Result***

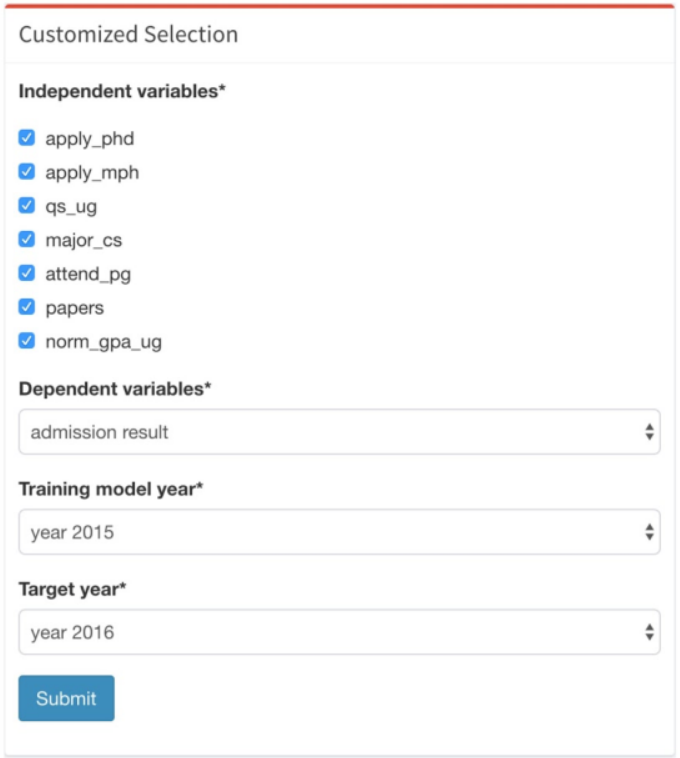
Applicant List

Show 10 entries Search:

ID	Predicted Result	Rank	Admit Probability	apply_phd	apply_mph	qs_ug	major_cs	attend_pg	papers	norm_gpa_ug
Y16A05	admit	1	0.383	phd	-	15	CS	-	0	0.9125
Y16K07	admit	2	0.373	phd	-	75	CS	-	0	0.96
Y16K05	admit	3	0.373	phd	-	75	CS	-	0	0.9425
Y16K02	admit	4	0.373	phd	-	75	CS	-	0	0.94
Y16M05	admit	5	0.372	phd	-	75	CS	-	0	0.9075
Y16K10	admit	6	0.372	phd	-	75	CS	-	0	0.9
Y16L07	admit	7	0.372	phd	-	75	CS	-	0	0.9
Y16K03	admit	8	0.371	phd	-	75	CS	-	0	0.885
Y16K06	admit	9	0.371	phd	-	75	CS	-	0	0.875
Y16G01	admit	10	0.37	phd	-	75	CS	-	0	0.8375

- Rank by probability
- Prediction / Recommendation

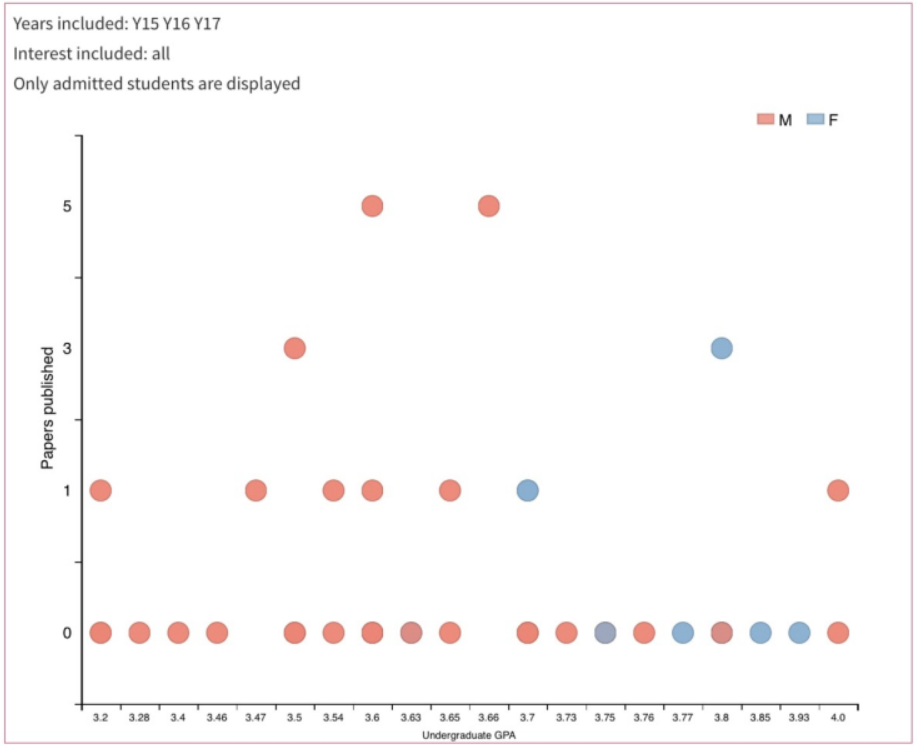
Logistic Regression *Customization*



CUSTOMIZE

- Independent variables
- Dependent variable
- Applicant data for training model
- Applicant data for prediction

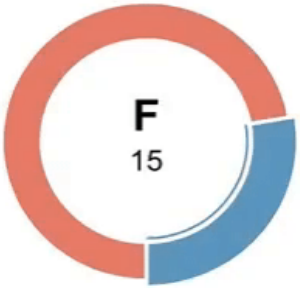
IMPLICATIONS



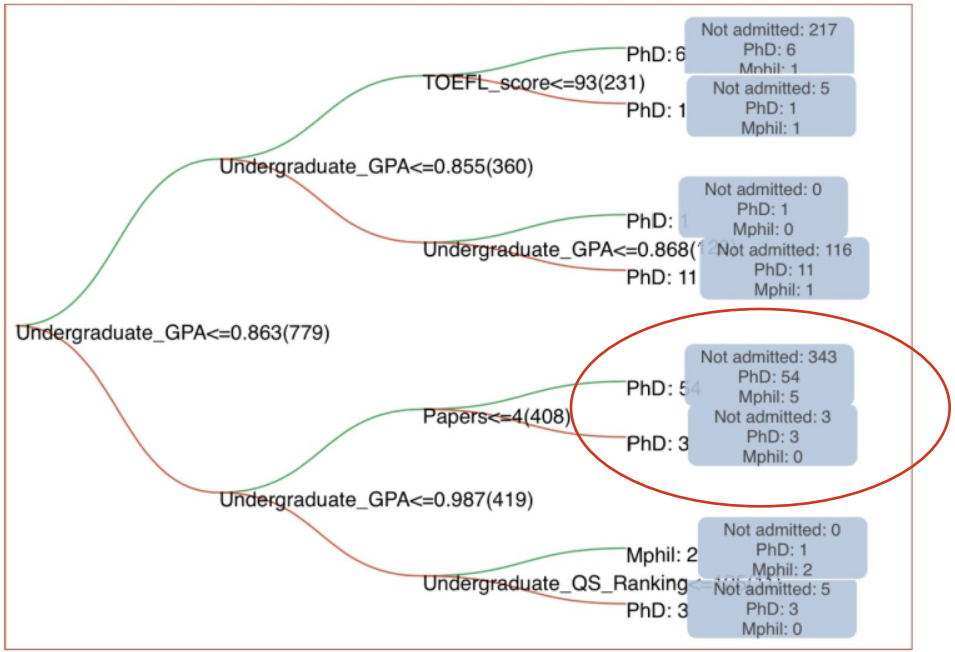
In 2015 - 2017

More than 70% of admitted students are

Perform generally better in terms of GPA and number of paper published



IMPLICATIONS



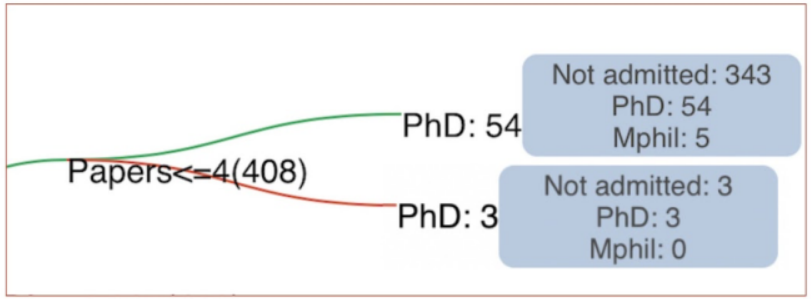
In 2014 - 2017

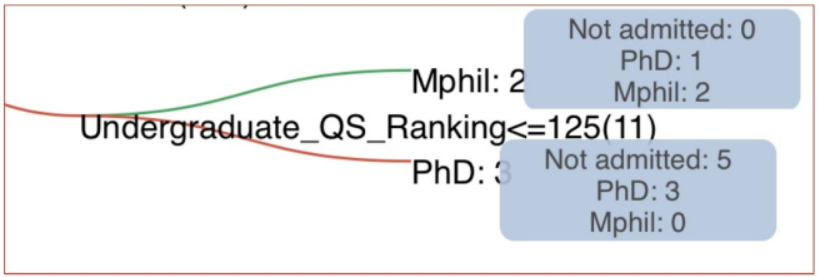
Similar Decision Rules for **PhD** and **MPhil**

0.86 < Undergraduate GPA <=0.98
Papers <=4

MPhil students have published less paper but come from better universities

IMPLICATIONS





In 2014 - 2017

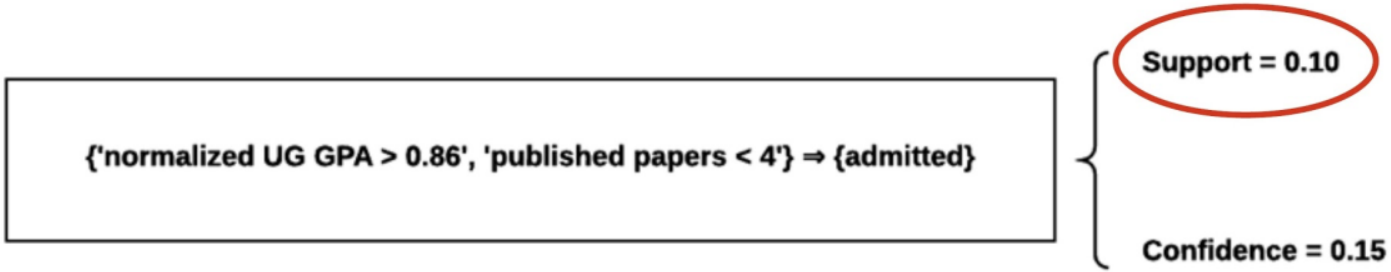
Similar Decision Rules for **PhD** and **MPhil**

0.86 < Undergraduate GPA <=0.98
Papers <=4

MPhil students have published less paper but come from better universities

IMPLICATIONS

Association Rule, 2015 - 2016 data

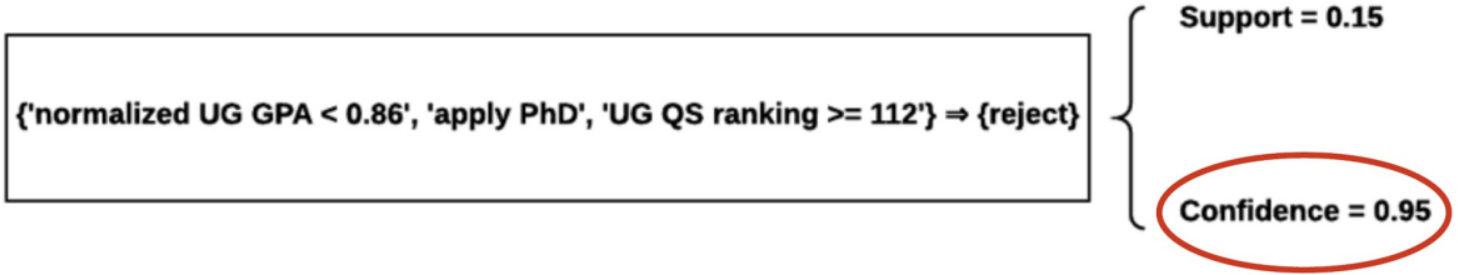


Conclusion:

Publishing more papers is not necessarily an advantage.

IMPLICATIONS

Association Rule, 2015 - 2016 data



Conclusion:

Give lower priority to applicants applying for PhD, coming for a UG with ranking higher than 111 and has GPA smaller than 0.86.

IMPLICATIONS

Logistic Regression, 2015 data

$$rej_i = 0.52 + 0.342apply\_phd_i + 1.49apply\_mph_i + 0.001qs\_ug_i - 0.299major\_cs_i$$

$$+ 0.737attend\_pg_i + 0.036papers_i - 0.1norm\_gpa\_ug_i + \in_i$$

IMPLICATIONS

Logistic Regression, 2015 data

$$rej_i = 0.52 + 0.342apply\_phd_i + 1.49apply\_mph_i + 0.001qs\_ug_i - 0.299major\_cs_i + 0.737attend\_pg_i + 0.036papers_i - 0.1norm\_gpa\_ug_i + \in_i$$

- MPhil applicants are more likely to be rejected

IMPLICATIONS

Logistic Regression, 2015 data

$$rej_i = 0.52 + 0.342apply\_phd_i + 1.49apply\_mph_i + 0.001qs\_ug_i - 0.299major\_cs_i$$
$$+ 0.737attend\_pg_i + 0.036papers_i - 0.1norm\_gpa\_ug_i + \in_i$$

- attend PG before, more likely to be rejected

IMPLICATIONS

Logistic Regression, 2015 data

$$rej_i = 0.52 + 0.342apply\_phd_i + 1.49apply\_mph_i + 0.001qs\_ug_i - 0.299major\_cs_i$$
$$+ 0.737attend\_pg_i + 0.036papers_i - 0.1norm\_gpa\_ug_i + \in_i$$

- more paper published, more likely to be rejected

IMPLICATIONS

Recommended Applicants

Show 10 entries Search:

ID	Predicted Result	Rank	Admit Probability	apply_phd	apply_mph	qs_ug	major_cs	attend_pg	papers	norm_gpa_ug
Y16A05	ad	1	0.381	phd	-	15	CS	-	0	0.9125
Y16K07	ad	2	0.371	phd	-	75	CS	-	0	0.96
Y16K05	ad	3	0.371	phd	-	75	CS	-	0	0.9425
Y16K02	ad	4	0.371	phd	-	75	CS	-	0	0.94
Y16M05	ad	5	0.37	phd	-	75	CS	-	0	0.9075
Y16K10	ad	6	0.37	phd	-	75	CS	-	0	0.9
Y16L07	ad	7	0.37	phd	-	75	CS	-	0	0.9
Y16K03	ad	8	0.369	phd	-	75	CS	-	0	0.885
Y16K06	ad	9	0.369	phd	-	75	CS	-	0	0.875
Y16G01	ad	10	0.368	phd	-	75	CS	-	0	0.8375

- MPhil applicants are more likely to be rejected
- attend PG before, more likely to be rejected
- more paper published, more likely to be rejected

CONCLUSION

- Educational Data Mining - HKUEDM
- Three Layers
  - Data Visualization
  - Data Analysis
    - Decision Tree
    - Association Rule
  - Prediction - Logistic Regression
- Implications from experiments
- Any Question ?

Reference

[1] Romero C, Ventura S. Educational Data Mining: A Review of the State of the Art. IEEE Trans. Syst., Man, Cybern. C. 2010; 40: 601-618.

[2] Mashat AF, M.fouad M, Yu PS, Gharib TF. Discovery of Association Rules from University Admission System Data. IJMECS International Journal of Modern Education and Computer Science. 2013; 5:1-7.

[3] Feng S, Zhou S, Liu Y. Research on Data Mining in University Admissions Decision-making. International Journal of Advancements in Computing Technology IJACT. 2011; 3: 176–186.

[4] S. Y.-W. B.-A. R. Fong S, "Applying a hybrid model of neural network and decision tree classifier for predicting university admission," 2009 7th International Conference on Information, Communications and Signal Processing (ICICS), 2009. [5] Django, https://www.djangoproject.com/start/overview/.

[6] QS World University Rankings by Subject 2016 - Computer Science & Information Systems, https://www.topuniversities.com/university-rankings/university-subject-rankings/2016/computer-science-information-systems.

[7] Apriori API, https://pypi.python.org/pypi/apyori/1.1.1.

[8] scikit-learn Logistic Regression Model, http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

Thank you !

Visit us at

http://i.cs.hku.hk/fyp/2016/fyp16019/

Contact

Johnson - wuyouk@connect.hku.hk

Carrie - carriex@connect.hku