X Modal
X Cultural
X Lingual
X Domain
X Site
Global OER Network

Grant Agreement Number: 761758
Project Acronym: X5GON
Project title: X5gon: Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network
Project Date: 2017-09-01 to 2020-08-31
Project Duration: 36 months
Document Title: Report on selected and evaluated quality assurance models
Author(s): Sahan Bulathwela, Emine Yilmaz and John Shawe-Taylor
Contributing partners: UCL, JSI, K4A, UPV
Date: 31/08/2018
Approved by:
Type: Report
Status: Final
Contact: m.bulathwela@ucl.ac.uk, emine.yilmaz@ucl.ac.uk, j.shawe-taylor@ucl.ac.uk

Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Revision

Date	Lead Author(s)	Comments
08/08/2018	Sahan Bulathwela	Initial Draft
20/08/2018	Sahan Bulathwela, Emine Yilmaz	Updates from the feedback from Emine Yilmaz
21/08/2018	Sahan Bulathwela, John Shawe-Taylor	Updates from the feedback from John Shawe-Taylor
28/08/2018	Sahan Bulathwela, Erik Novak	Updates from the feedback from Erik Novak
29/08/2018	Sahan Bulathwela	Adding the list of figures, list of tables, abstract and appendix
31/08/2018	Sahan Bulathwela	Final Version

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Abbreviation	Meaning
OER	*Open Educational Resource*
VLN	*videolectures.net*
SVM	*Support Vector Machine*
MTL	*Multi-task Learning*
NLP	*Natural Language Processing*
IJS	*Institut "Jožef Stefan"*
JSI	*Institut "Jožef Stefan"*
API	*Application Programming Interface*
RMTL	*An R Library for Multi-task Learning*
RR	*Ridge Regression*
RMSE	*Root Mean Squared Error*

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

X5GON (Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network) is a Horizon 2020 collaborative project aiming at providing the next generation network for learning with Open Educational Resources. The goal of this project is to leverage Information Technology and Artificial Intelligence to develop a learning resource network that an provide high quality, personalised learning pathways to learners by recommending open educational resources from multiple repositories.

This is the second deliverable of the work package relating to learning rich content representations (WP1) reporting on the evaluation and selection of quality models to assess the content quality of OERs automatically at scale. This report outlines the comparison between the 3 main models that are described in deliverable D1.1 to select the most suitable machine learning model for quality assurance.

Based on the final comparison of results, the SVM model that uses pairwise comparison technique performed best on the test data with 71% classification accuracy. Although the MTL model performed as well, the SVM model is more desirable as the final choice as it achieves similar or superior test data performance with a much simpler model. Further analysis of the weight coefficients of the SVM model also shows that the patterns learned by the model are sensible.

Taking all these factors into account, Ranking SVM model is the most suitable quality assurance model due to its high performance and interpretability.

1. INTRODUCTION

Assessing the content quality of open education resources automatically, at scale is an essential part of deriving accurate representation for building systems for high quality personalization of learning trajectories.

1.1. Quality Assurance in X5GON

In the context of Open Educational Resources (OER), quality of a resource can be defined as an attribute of content that encourages / discourages interaction between the learner and the resource. When recommending effective learning materials to learners in X5GON, it is very important that we recommend them resources that help them expand their horizons. But learning to happen, X5GON should have reasonable confidence that the learners are likely to interact with the recommendations. This is where the quality assurance front plays a big role.

In the earlier stages of the project, quality assurance models are developed to automatically identify high quality content from low quality content. Quality assurance models help in three main ways.

1. Allows automatic identification of high quality vs. low quality education resources when X5GON integrates with new repositories
2. Allows ranking lectures and comparing between them to automatically rank them based on content quality
3. In the long term, leverage personalization by capturing patterns about quality features different users prefer when consuming OERs.

In this report, we expand the model evaluation and model selection process involved with Quality assurance models. We explain the various models that were developed. Then we explain how the models were evaluated and the criteria under which the final quality assurance model was selected. This document will focus on the model evaluation and selection steps of the quality model development process. For a more descriptive overview of the literature survey and the model formulation steps, we direct to section 4: “About quality assurance in educational content” in deliverable D1.1.

1.2. Main Contributions

The main contributions made in the first 12 months of the project are as follows:

1. **Literature Review on Quality Assurance:** Given that quality of education resources is an ill-defined topic, a thorough literature survey was done exploring to different knowledge domains to understand what a sensible definition of quality of content is.
2. **Data Collection:** Obtaining required authorisations and harvesting data from the data sources available to use through our partners. As the dataset was fresh several cycles of data cleaning and sanity checking was carried out to ensure the reliability of data.
3. **Data pre-processing and deriving a quality label based on user engagement:** Based on the findings from the literature survey, data and the tools available, a set of features and labels were derived from the raw dataset.
4. **Developing several models for predicting quality of educational resources:** Developed several supervised learning models to predict the quality of educational material.

5. **Rigorous evaluation and selection of a potential model:** Evaluated both regression and classification models developed. Due to the differences between evaluation metrics used to evaluate different models, we formulated a strategy to compare the models fairly. Based on the evaluation, a sensible model was selected for deployment.

1.3. Document Overview

In ***section 2***, we describe the datasets that were used during this study. The top 1,000 lectures dataset and the full VLN dataset is explained briefly.

Continuing to ***section 3***, we discuss the models developed using the datasets described in the earlier section.

Final methodology used in evaluating multiple models is outlined in ***section 4*** with multiple potential metrics that are available. Then the final set of evaluation metrics selected is outlined in ***section 5*** with relevant rationale.

Section 6 describes the results obtained from model evaluation phase. From the evaluation results, we extend to ***section 7*** where the model weights and misclassifications are analysed in detailed. Finally, outline the future research avenues in section 7.

In ***section 8***, we sum up the study by deriving conclusions.

2. DATASETS

Videolectures.net repository is the main source of data available for the first phase of X5gon. Data was extracted from the main repository in two stages.

2.1. Background context

Since the inception of the project, a lot of things had to be setup to pave way to the quality models to realise. The most important task in the initial stage is to leverage relevant data to train the machine learning models on. Most efforts in the first phase of the project was put towards granting access to relevant data from www.videolectures.net (VLN) that was available to the project. Once the required authorizations were obtained, a significant amount of time was spent on downloading raw data, understanding the data, cleaning the data and compiling it into a usable dataset to be used to develop machine learning models for quality assessment. Data is described in detail in section 6.1 in deliverable D1.1.

As this is a brand-new dataset that has not been analysed by the academic community before, rigorous measures were taken to sanity check every step of data processing with no prior assumptions about the correctness of data.

It is also fair to draw attention to the fact that VLN was the primary and only source of data available in this phase. Careful evaluations should be done once new data sources are available to ensure the generalizability of the developed models to any educational resource.

2.2. Top 1,000 lectures dataset

Initially, a dataset that comprises of the 1,000 most popular lectures in VLN repository was extracted with lecture-related data such as duration of the lecture, title, description, hotness score etc.

$$Hotness = \frac{Number\ of\ views}{Number\ of\ days\ since\ publication^2} \tag{1}$$

The quality in this dataset is measured using “hotness score” computed using (1). This dataset consists of the lectures that has the 1000 highest hotness scores precomputed on the queried date.

Unfortunately, we were unable to find any literature that uses this formula to represent quality of content. However, this is the metric currently used by VLN repository for ranking lectures.

2.3. Full videolectures.net dataset

This is the raw data relating to all the lectures in VLN repository. This dataset contains lecture data until April 2018 and was downloaded with the help of JSI from VLN repository using the data API. This dataset consists of 25,697 lectures with details about their authors, authors’ affiliations, user engagement related data etc... In terms of potential target variables, average star rating, user engagement data and hotness scores are available for most lectures in this dataset. A detailed description about the variety and volume of this dataset can be found in section 6.1 in deliverable D1.1.

3. DEVELOPED MODELS

Several regression and classification models were developed based on the two datasets mentioned in section 2. In this section we will focus on a subset of those models.

3.1. QUALITY MODELS WITH TOP 1,000 LECTURES DATASET

Based on the initial dataset outlined in subsection 2.2, a linear classification model was developed to classify the quality of educational resources. In summary, the dataset was separated into 4 main classes based on the hotness score quartiles. Then a logistic regression model was used. Unfortunately, we conclude that this dataset is not representative in terms of learning quality dynamics of educational resources. The reasons for this conclusion are outlined below.

1. **Over representative of higher quality resources:** The dataset contains the 1,000 most popular lectures in VLN repository. This means that there is minimal representation of bad quality lectures in this dataset.
2. **Subject Bias in the dataset:** VLN repository has a strong presence of computer science, machine learning and deep learning related videos. Due to recent emergence of Deep Learning and other machine learning sub-domains, majority of the most popular lectures are machine learning related lectures. Therefore, the top 1,000 datasets are heavily biased towards machine learning related content.
3. **Hotness doesn't completely represent quality:** Hotness score is more of a popularity indicator than a quality indicator. As YouTube found out, keeping up clicks (popularity) doesn't really mean that the content possesses good quality *[Meyerson (2012)]*. Therefore, hotness score may not be the best target variable to measure quality of content.
4. **Very small dataset:** The dataset only has 1,000 datapoint. There are obvious better alternatives such as the "full" dataset from VLN repository.

However, we include the results from the logistic regression model fit to the top 1,000 lectures dataset in *Appendix A1*.

3.2. QUALITY MODELS FROM THE MAIN VLN REPOSITORY DATASET

During the model development phase, three main models were developed.

1. Ridge Regression Model (RR)
2. Pairwise Classification model using Support Vector Machines (Ranking SVM)
3. Pairwise Multi task classification model using trace norm (Trace MTL)

To get a more detailed explanation about RR, Ranking SVM and Trace MTL models, subsections *5.3.1, 5.3.2 and 5.3.3 in deliverable D1.1* can be referred to respectively.

As Quality Assurance model derivation is treated as a supervised learning problem, A series of features:

- Document Entropy
- Easiness
- Fraction of Complex Words
- Fraction of Silent Words
- Fraction Stopword Coverage
- Fraction Stopword Presence
- Published Date Epoch Days
- Title Word Count

- Word Count

Were used. Please refer to subsection *6.3.1: in deliverable D1.1* for a detailed account of how exactly the features are computed.

3.2.1. Target Variables for Quality

The primary label used as the target variable is "**median engagement rate**". We first compute the fraction watch time of each user for the lecture as engagement rate. Then we select the median amongst the engagement rates for each lecture and use that value as the label for the lecture.

- When the median engagement rate is close to 0.0, this means that the users watch a very small fraction of the video lecture.
- When the media engagement rate is close to 1.0 or higher[1], this means that the users watch most of the video

Initially, there were 3 potential target variables available in VLN Dataset.

1. Average Star Rating of the lecture
2. Hotness Score
3. Engagement related data

According to literature, explicit labels such as star ratings are very powerful signals when measuring user perception towards a piece of content *[Amatriain (2009, 2012)]*. A lot of services in the Internet uses stars, likes *[Kincaid (2009)]* and other forms of explicit feedback to capture user perception. Unfortunately, explicit feedback is very expensive to acquire making it a scarce signal. Also, explicit feedback may not always be as objective and representative of what is being measured. Psychological theories such as Theory of Planned Action *[Ajzen (1991)]* suggest that explicit signals usually represent personal attitudes and biases but also heavily influenced by what people perceive other people around them to prefer as well, not their sole personal preferences.

Hotness score on the other hand, represents popularity of a lecture rather than quality as equation (1) in section 2.2 suggests. In addition, we were unable to find formula (1) in section 2.2 used in any literature applied to representing quality. Therefore, we conclude hotness is an unsuitable target variable.

Due to these reasons, video engagement is the most promising and readily available target variable available. There is numerous works that also video watch time and implicit variables being suitable to measure engagement *[Meyerson (2012)]*.

For exact details about how the median engagement labels are computed, please refer to subsection *6.3.2: in deliverable D1.1*.

[1] When a user replays component of a video repeatedly in the same session, the total watch time exceeds the duration of the video. Hence the engagement rate can be greater than 1.0. As the median is used as the centre, it is not sensitive to massive engagement rates due to replaying.

4. METHODOLOGY

Evaluation gives the confidence required to conclude that any given model is going to behave as expected. It also gives the framework to objectively compare between multiple alternatives.

4.1. DESIRED SOLUTION

The ideal solution would desirably,

1. **Have good generalization performance:** Have good performance on held out data and robustly perform well on new educational content. The ideal model wouldn't overfit to the training data.
2. **Be easily interpretable:** It is ideal to have a simple model that is interpretable. As the task at hand is to predict the quality of educational resources, it is desirable to have a model that is explainable as quality in educational material is a sensitive topic.
3. **Be sensible:** The suitable model should show evidence that it captures patterns that define quality and not capturing something else. Being able to validate this is very powerful.

4.2. EVALUATION METRICS

In this section, we discuss a few evaluation metrics that are available for us to evaluate the models we are building.

For the detailed definition of all the evaluation metrics mentioned in subsections 4.2.1 and 4.2.2, please refer to *Appendix A2: Definitions of Evaluation Metrics.*

4.2.1. Evaluating regression models

In regression model scenario, Metrics such as Root Mean Square Error (RMSE) and Coefficient of determination (R^2) is used to measure the predictive power of models. RMSE and R^2 are geared towards measuring the accuracy of predicting a real value (when $y \in \mathbb{R}$).

But in this scenario, the objective is to use a regression model to predict values in such a way that the order of observations is accurate. Therefore, the exact deviation of the prediction from the true value is not as important if the global order is preserved. Spearman Rank Correlation and Kendall Tau Rank Correlation are two of the main evaluation metrics that are used to evaluate if two ranked lists have similar order.

4.2.2. Pairwise Classification Models

Pairwise preference is an approach used to convert a ranking problem into a classification problem *[Joachims (2002), Herbrich et al (1998)]*. By converting the problem in to a pairwise preference problem, it is possible to express the ranking landscape in more detail using N^2 observations.

Precision, Recall and Accuracy scores are a few metrics that are widely used when evaluating classification models.

4.2.3. Sanity checking with Weight Coefficients

Sanity checking the derived model is an essential part of model selection. This allows us to understand if the model has learned something that is contradictory to our beliefs. Analysing the weight coefficients is a very sensible way to do this. In a scenario like this where there is not much prior knowledge about what features drive quality of content, weight coefficient analysis is a great way to investigate both interpretability and sensibility desired in the solution (section 4.1).

4.3. COMPARING MODELS

We are considering both regression and classification models as the future quality assurance model. Conventionally, these models are evaluated differently. We devise classification accuracy as the global evaluation metric used to evaluate between all models. Section 5.2.3 explains in detail how the evaluation is done.

5. EVALUATION AND MODEL SELECTION CRITERIA

In this section, we outline the overall evaluation and model selection criteria devised. We first define the test data and then the evaluation criteria.

5.1. TEST DATA

As outlined in *section 7 in deliverable D1.1*, we leave 30% of the labelled data as testing data according to Table 1. This data will not be used during the training phase. The data is partitioned between training and test sets using stratified sampling on the field subject. Therefore, both training and test set will have similar proportions of examples from every field category[2] (Biology, Computer Science etc...).

Dataset	Proportion (%)	Frequency (Raw)	Frequency (Pairwise)
Training Data	70.0	3,619	3,970,218
Testing Data	***30.0***	***1,562***	***729,422***
Total	100.0	5,181	4,699,640

Table 1: The train test split for raw and pairwise datasets

When pairwise comparisons are generated, we generate the pairwise comparisons after we split the raw lecture data in to train and test sets. This guarantees that both comparison data points from the same lecture pair will be restricted to either training or test set, but not both.

For a more detailed explanation of data generation, please refer to sections 6 and 7 in deliverable D1.1.

5.2. FINAL EVALUATION STRATEGY

We employ a subset of metrics discussed in section 4.2 as our final evaluation strategy.

5.2.1. Model evaluation with test data

When evaluating the regression models, we use RMSE and Spearman's rank correlation. We use RMSE because it tells how deviated the predications are from the true label (median engagement rate).

However, as the main objective of the task is to rank lectures, a rank correlation metric is necessary to measure the overall accuracy of the model in terms of ranking. We use Spearman's rank correlation for this. In this scenario, it is highly unlikely that the global ranks of lectures based on their engagement rate would have ties. Therefore, Spearman correlation is good enough.

We use Classification accuracy to measure pairwise preference performance. Classification accuracy metric has a natural interpretation when used to measure performance of pairwise preference model. It represents how likely the model would correctly predict a preference outcome when comparing any two lectures from the same field. However, it is hard to define precision and recall for pairwise comparison.

We analyse the weight coefficients from the models to further sanity check if the models make intuitive sense.

[2] For a full list of field categories, refer to Table 1 in deliverable D1.1

5.2.3. Comparing and selecting the final model

While we develop both regression based and pairwise preference classification based models, we need a method to compare all these models in same grounds. From above section, it is seen that we use very different quantitative metrics to evaluate them. However, this will hinder us from comparing between the regression models and classification models.

5.2.3.1. Comparing regression model with the classification models

In this scenario, we solve a ranking problem· Regression derives a global ranking whereas the preference models derive a relative ranking system.

Global to Pairwise preference: Convert the global ranks to a pairwise preference dataset. As shown by (8), it is possible to convert a global rank to a unique pairwise representation.

$$\begin{aligned}
&global\ ranking:\\
&a > b > c\\
&\\
&can\ be\ converted\ to:\\
&a > b = True\\
&b < a = False\\
&b > c = True\\
&c < b = False\\
&a > c = True\\
&c < a = False
\end{aligned} \tag{8}$$

Pairwise preference to global: It is also possible to convert a pairwise preference ranking to a global rank *[Fürnkranz & Hüllermeier (2010)]*. But, there is more than one unique solution for this problem. Especially when there is misclassification error associated with the preference classifier.

The most reliable comparison is to convert the solution from regression model to a pairwise preference dataset and compare it with the classification models. This conversion will always have a unique solution that can be compared against the other results. Now that all predictions are in pairwise format, Classification accuracy can be used to compare between the models.

5.3. SELECTION CRITERIA

As discussed in *section 4.1*, the focus is on good generalization performance. Therefore, the highest priority from the metrics is given to test data performance. Furthermore, we increase the reliability of the selected model by eyeballing weight coefficients.

6. EVALUATION RESULTS

In this section, we report the results from the evaluations. Then we move forward to select the most suitable model for quality assurance in educational resources.

6.1. Ridge Regression model

RMSE and Spearman correlation coefficient (*Spearman R*) has been used to evaluate the regression results obtained by Ridge Regression. Table 2 summarises the evaluation results from the ridge regression model trained with the data.

Evaluation Metric`	Training Data	Test Data
RMSE Training data	0.1907	0.1838
Spearman R (p-value[10])	0.5638 (7.63e-303)	0.5814 (6.01e-142)

Table 2: *Model evaluation results from Ridge Regression*

6.2. Pairwise Classification Models

In the classification setting, classification accuracy is the metric used. Table 3 below summarises the classification accuracy score obtained for both SVM (Ranking SVM) and Multitask Classification using Trace norm (Trace-Norm MTL).

Classification Accuracy	Training Data	Test Data
Ranking SVM	0.7191	0.7121
Trace-Norm MTL	0.7210	0.7105

Table 3: *Classification Accuracy of Ranking SVM and Trace Norm MTL models*

6.3. OVERALL COMPARISON OF MODELS

We use classification accuracy as the metric that is generalizable to all the models developed (explained in *section 5.2.3*). Table 4 summarises the results from the classification accuracy results from the three models under investigation.

Classification Accuracy	Training Data	Test Data
Ridge Regression	0.7120	0.7115
Ranking SVM	**0.7191**	**0.7121**
Trace-Norm MTL	0.7210	0.7105

Table 4: *Final Comparison of all three models (i) Ridge Regression, (ii) Ranking SVM, and (iii) Trace-Norm MTL*

6.4. Observations and Conclusions

From the results in section 6.1, 6.2 and 6.3 we can draw a few notable observations.

1. All the models are very robust when it comes to overfitting. It is seen that the disparity between the training error and testing error in all the models are very small. This means that the models are very well defended against overfitting to training data.
2. According to table 4, test set performance of all the models are quite similar. They all perform at around 71% classification accuracy.
3. Ranking SVM has the highest classification accuracy of 71.21%
4. Multitask learning approach doesn’t seem to give a significant advantage in performance over the linear models although it has more learnable parameters.
5. Multi task learner also shows tendencies of overfitting according to *tables 4 and 5.*

6.5. Conclusions

The main conclusion from the above observations is that the Ranking SVM has the best test set performance and hence the strongest candidate for the potential quality assurance model.

Another major conclusion is that the Trace norm regularised multi task learning model hasn't improved results in comparison to the linear models. This suggests that the trace-norm regularization does not help the model to learn better than linear models although there is more freedom to learn individual weights for different field categories.

From the evaluation results, linear models seem to perform well while preserving highest degrees of interpretability. It is helpful to eyeball and sanity check the linear models further before selecting the final model used for lecture quality assurance.

7. ANALYSING THE MODEL

In this section, we focus on sanity checking the selected linear models to further confide that the models are detecting patterns that are attributable to quality of content.

Furthermore, we explore into the misclassifications in the model to understand what factors drive the prediction mistakes.

7.1. WEIGHT COEFFICIENT ANALYSIS

Weight coefficient analysis is a very useful technique to sanity check a learned model. A lot of studies in Econometrics and Social Science *[Eamonn et al (2002), Weerahewa et al (2012)]* and Natural Language related problems *[Liang et al (2018)]* use weight coefficient analysis to interpret the actual patterns driving the data generation process.

The weight coefficients of the two linear models, Ridge Regression (RR) and Ranking SVM (SVM) are outlined in Table 5. Please refer to subsection 3.2 for detailed account of the features in the models.

Feature	RR	SVM	Interpretation
Document Entropy	0.04720	0.27343	Preferred when the lecture takes about several topics rather than one focussed topic
Easiness	-0.10334	- 0.54592	Advance language is preferred
Fraction Complex Words	-0.04053	- 0.30920	Shorter, simpler words are preferred
Fraction Silent Words	0.02951	0.10986	Having pauses during the lecture keeps learners engaged
Fraction Stopword Coverage	-0.16509	- 0.74089	Prefer less presence of stopwords -> advance language
Fraction Stopword Presence	0.00511	- 0.04817	Two models contradict, but the weight coefficient is very small
Published Date Epoch Days	0.01345	0.06243	Fresh content is preferred over older content
Title Word Count	0.00303	0.01735	Lectures with longer titles tend to increase engagement
Word Count	0.03981	0.15726	Longer lectures are preferred over shorter lectures.

***Table 5:** Weight Coefficient Analysis of the linear models*

It is evident from Table 5 that both linear models learn very similar patterns. The only contradiction occurs in one feature (Fraction Stopword Presence). But, this is the feature has got a significantly small weight coefficient which has a significantly smaller contribution to the final label in the decision function. Overall, the interpretations of the respective weight coefficients are sensible.

After we have analysed the weight coefficients, we confidently select the Raking SVM model as the final model as it shows good test set accuracy while characterising sensible weight coefficients.

7.2. Diagnosing Misclassifications

Understanding sources of error is an essential part of improving the model at hand. Therefore, we focus on the subset of observations that are misclassified by the SVM model. First, we identify what potential factors are contributing to these mistakes.

7.2.1. Field Category Biases

One main potential factor for error would be if the model was significantly underperforming on a subset/ cluster of subjects. We can investigate this by looking into the accuracy of each subject in the dataset.

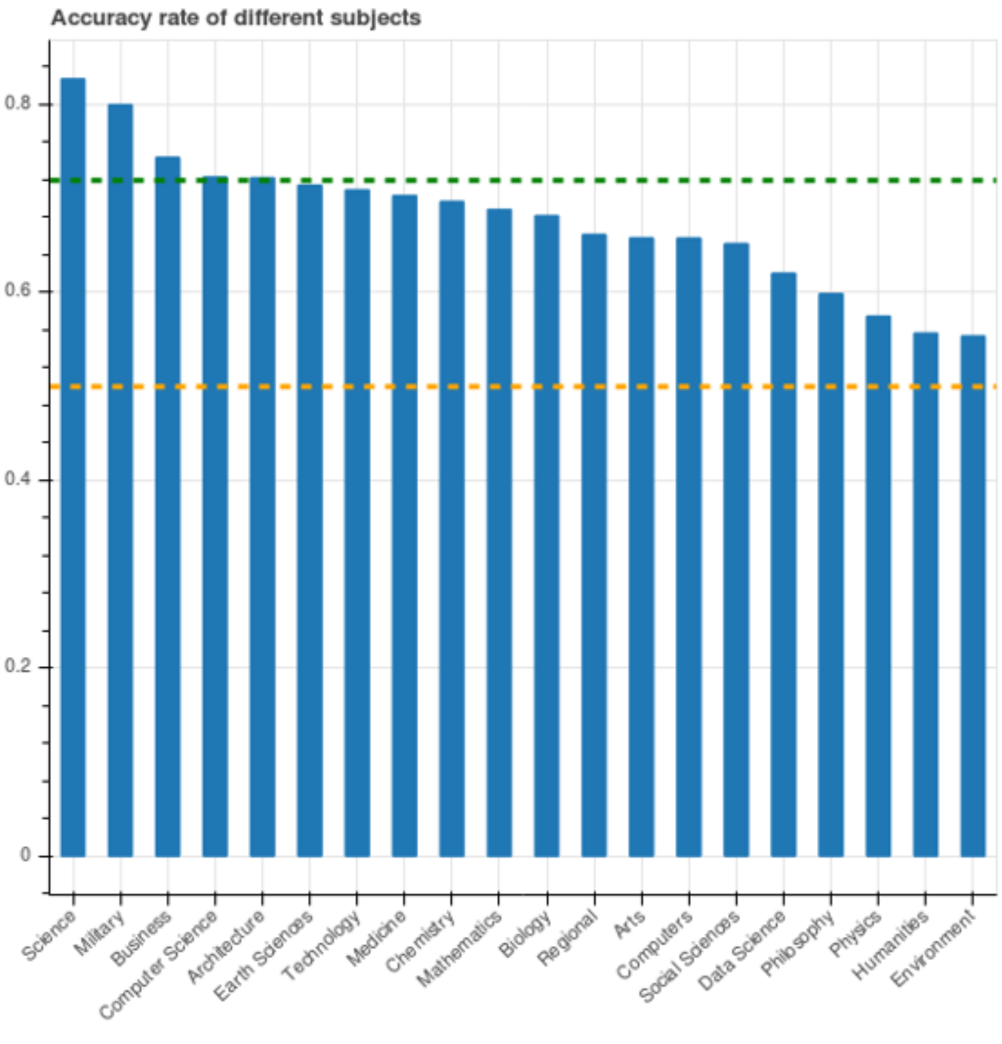


Figure 1: *Accuracy of model by Field category*

Figure 1 outlines the accuracy rate by every field category in the dataset ordered from highest accuracy to lowest accuracy. From the figure, it is observable that there is no obvious trend of certain subjects systematically underperforming (e.g.: Science Subjects vs. Arts subjects and etc...). Ordering the subjects by accuracy score doesn't emerge any dominant patterns in subject-wise accuracy rates.

Furthermore, it is seen that all the field categories are performing better than random (random is yellow dotted line at 0.5 accuracy). Most field categories perform close to 0.71 accuracy (shown by green dotted line) leading to overall accuracy rate of 0.71.

7.2.2. Difficult examples

Another potential reason for misclassification is if a subset of examples is systematically difficult to classify. As explained in deliverable D1.1, we convert the real valued gap between engagement rates ($y_{l1} - y_{l2} \in \mathbb{R}$) to a discrete value which is {True, False} according to (13) in subsection 6.3.3 in deliverable D1.1.

Figure 2 below shows the normalised histograms of correctly predicted (green) and wrongly predicted (red) test set data where the horizontal axis is the difference of engagement rates between lecture 1 and lecture 2 ($y_{l1} - y_{l2} \in \mathbb{R}$).

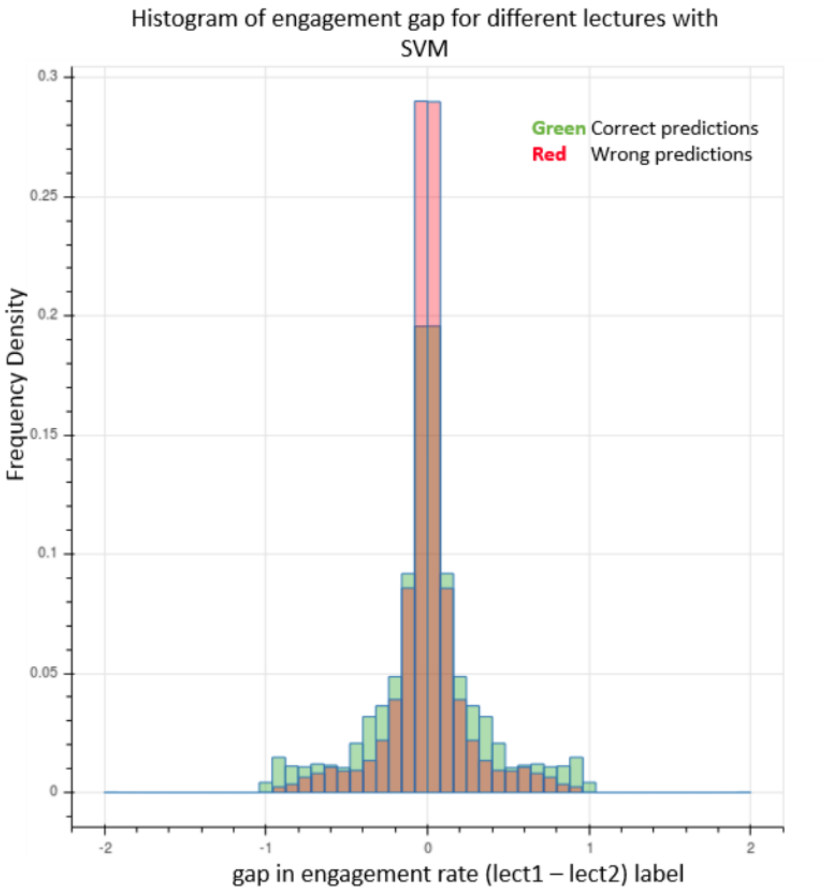


***Figure 2:** Histogram of Real Gap between engagement rates to frequency density of test data for correct (green) and wrong (red) predictions*

It is highly evident from histograms in Figure 2, that the misclassification rate of the model significantly increases when the actual engagement rate gap is very close to zero. According to the plot, the classifier does significantly better in correctly classifying the superior lecture when the difference between engagement rates gets higher. Similarly, the misclassification rate increases significantly when the gap between the engagement rate is closer to zero.

This observation is further confirming that the model is learning exactly what is expected to be learned. It is true that the model is doing bad around 0. But the change in misclassification error when the target value is getting farther from 0 suggests that the model is learning to attribute the difference of features to the difference in engagement rate.

7.3. Future Work

In the first 12 months, we focussed on building well understood models that perform very well (section 4.1). The models we have developed so far can be considered as quite basic. However, according to the evaluations, these models prove to be quite powerful in terms of performance and interpretability.

For quality assurance in educational materials, we have built a strong foundation by understanding the patterns that govern quality (Section 7.1) while gaining valuable insight into main sources of error (section 7.2). Based on this knowledge we have already started exploring avenues on how to build on top of the knowledge we have acquired to build more performant models.

Because we have a clear understanding of how linear models perform on this problem, we also see potential in introducing more advanced neural models to improve performance. Results in Table 4 suggest that ranking based models may be more suitable for this problem. Hense, models such as RankNet *[Burges et al (2005)]*, LambdaRank *[Burges (2005)]* and LambdaMART *[Burges et al (2006)]* that have shown to outperform RankingSVM *[Joachims (2002)]* can be used make to performance gains very easily.

Results in section 7.2 opened a whole new avenue of thought for us. Based on the results, we are exploring if we can treat the labels as soft labels based on the gap between engagement levels of lectures by assigning uncertainty levels to true labels.

Eg: if | lecture_{l1} – lecture_{l2} | > | lecture_{l3} - lecture_{l4} |

We can be more certain that l1 > l2 than l3 > l4

In other words, *P (l1 > l2) > P (l3 > l4)*

One option is to use the user sessions for the lectures separately to compute a t-statistic for the difference of engagement. It is also possible to use a more sophisticated model like TrueSkill model *[Herbrich et al (2006)]*.

While we progress on quality assurance front, we can adapt the representations more towards personalization of content. We think that wikification *[Brank et al (2017)]* of educational material is an obvious way forward to capture knowledge contained in an OER. We can use these knowledge representations to model users with an adaptation of Bayesian Knowledge Tracing Model that enables searching for documents based on learners' knowledge needs *[Sayed & Collins-Thompson (2017)]*. As the quality models are based on engagement of users with content (refer section 3.2.1), it is very likely that there are connections between the quality of a resource and the amount of information a learner absorbs from that resource.

In a nutshell, our current work has led us to start exploring various interesting facets of learning and building up to more advanced and powerful solutions to improve our work. We look forward to describing our results in the reports to come.

8. CONCLUSIONS

From sections 6 and 7, we have been able to gain great insights into the models that were developed for predicting quality of educational resources. We evaluate and compare between 3 main models during this study.

1. Ridge Regression (RR)
2. Pairwise Preference Classification using SVM (SVM)
3. Multi-task Learning Pairwise Classification using Trace-norm regularization (MTL)

8.1. OBSERVATIONS

Section 6, we use classification accuracy which has a natural interpretation that can be defined as the probability of the classifier correctly predicting the educational content with superior quality.

When comparing the 3 above mentioned models using classification accuracy, the result shows that all three models developed perform similarly with about 0.71 accuracy (Table 4). The classification accuracy is very similar between the simple task linear models (RR and SVM) and the multi-task learner (MTL) although MTL model has the capacity to learn separate sets of weight coefficients for individual fields.

Considering this information, it is fair to conclude that the linear models are suitable for this task over multi-task learners as same performance can be achieved using a far simpler model.

Further investigation into the weight coefficients of the linear models (RR and SVM) leads us to believe that they consistently learn very similar patterns (Table 5). Furthermore, the weight coefficients are sensible as well. This observation leads us to the conclusion that the SVM is the ideal model for quality evaluation as SVM has smaller generalization error against RR as per Table 4.

Once the model has been selected, we further make attempts to identify the factors contributing to misclassifications. By looking at the correlation between the real difference between lecture engagement rate vs. misclassification error, we can observe that the classifier is significantly bad in identifying the superior lecture when the difference between median engagement rates is very small.

This observation further confirms that the model is learning to classify the difference between the engagement rates between two lectures. Another main thought-provoking idea is to ponder on how important it is to correctly predict the better lecture when their engagement rates suggest that they are very close in terms of quality. One could argue that it is not that important to identify the more superior resource when they are not significantly different in terms of quality.

8.2. CONCLUSION

Under these observations, we can safely conclude that using the SVM model to predict the pairwise superiority between lectures is a very reliable way to enforce quality assurance. Furthermore, the model is quite robust against unseen examples making it highly generalizable.

REFERENCES

Eric Meyerson. 2012. Youtube now: Why we focus on watch time. (August 2012). Retrieved July 30, 2018 from http://youtubecreator.blogspot.com/2012/08/youtube-now-why-we-focus-on-watch-time.html

Xavier Amatriain. 2009. The Netflix Prize: lessons learned. TechnoCalifornia (29 September 2009). Retrieved July 23, 2018 from http://technocalifornia.blogspot.com/2009/09/netflix-prize-lessons-learned.html

Xavier Amatriain, 2012. Building industrial-scale real-world recommender systems. In Proceedings of the Sixth ACM Conference on Recommender Systems, RecSys ’12, pages 7–8, New York, NY, USA, 2012. ACM

Jason Kincaid. 2009. Facebook Activates "Like" Button; FriendFeed Tires Of Sincere Flattery (February 9, 2009). TechCrunch. AOL. Retrieved August 08, 2018 from https://techcrunch.com/2009/02/09/facebook-activates-like-button-friendfeed-tires-of-sincere-flattery/

Icek Ajzen. 1991. "The theory of planned behavior". *Organizational Behavior and Human Decision Processes.* 50 (2)*:* 179–211

Thorsten Joachims. 2002. Optimizing Search Engines Using Clickthrough Data. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’02), 133–142. DOI: https://doi.org/10.1145/775047.775067

Ralf Herbrich, Thore Graepel, Peter Bollmann-Sdorra, and Klaus Obermayer. 1998. Learning preference relations for information retrieval. In ICML-98 Workshop: text categorization and machine learning, 80–84.

Johannes Fürnkranz, Eyke Hüllermeier (eds). 2010. Preference Learning and Ranking by Pairwise Comparison. In: Fürnkranz J., Hüllermeier E. (eds) Preference Learning. Springer, Berlin, Heidelberg. DOI:https://doi.org/10.1007/978-3-642-14125-6_4

Ferguson Eamonn, James David, Madeley Laura. Factors associated with success in medical school: systematic review of the literature BMJ 2002; 324 :952

Jeevika Weerahewa, Sahan Bulathwela, Pradeep Silva, Kalyani Perera, 2012, An Analysis of Academic Performance of Undergraduates: Effects of Academic Vis-A-vis Non-Academic Factors, Peradeniya University Research Sessions (PURS), Vol. 17

Shangsong Liang, Emine Yilmaz and Evangelos Kanoulas. 2018. Collaboratively Tracking Interests for User Clustering in Streams of Short Texts. in *IEEE Transactions on Knowledge and Data Engineering*.

Christopher Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning* (ICML '05). ACM, New York, NY, USA, 89-96. DOI=http://dx.doi.org/10.1145/1102351.1102363

Christopher Burges. Ranking as Learning Structured Outputs. 2005. *Neural Information Processing, Systems workshop on Learning to Rank*, Eds. S. Agerwal, C. Cortes and R. Herbrich, 2005

Christopher Burges, Robert Ragno, Quoc V. Le. 2006. Learning to Rank with Non-Smooth Cost Functions. *Advances in Neural Information Processing Systems*, 2006.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkillTM: A Bayesian Skill Rating System. In Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS'06), 569–576. Retrieved August 30, 2018 from http://dl.acm.org/citation.cfm?id=2976456.2976528

Janez Brank, Gregor Leban, Marko Grobelnik. Annotating Documents with Relevant Wikipedia Concepts. 2017. *Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017)*, Ljubljana, Slovenia, 9 October 2017

Rohail Syed and Kevyn Collins-Thompson. 2017. Retrieval Algorithms Optimized for Human Learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '17). ACM, New York, NY, USA, 555-564. DOI: https://doi.org/10.1145/3077136.3080835

APPENDIX

A1: RESULTS FROM TOP 1000 LECTURES DATASET

Following section outlines a summary of the analysis done on top 1000 lectures dataset described in section 2.2.

A1.1. Introduction

Initially, a dataset that consists of 1,000 lectures in VLN was handed to be used for building quality assurance models. These lectures were ranked using an internal metric called “hotness score” explained in section 2.2.

A1.2. Dataset

As mentioned above, the dataset consisted of 1,000 lecture records. The English transcripts relating to the contents of the lectures was also available. Some additional fields related to the lectures such as lecture title, authors, lecture summary and description were available in the dataset.

A1.3. Target Labels

The lecture records in this dataset were the 1,000 top ranking lectures based on VLN repository’s internal ranking metric, hotness score. For further details on hotness score, you can refer to section 2.2.

Initially, we investigated the hotness score distribution and found out that it is highly skewed. It is evident that hotness score gets discounted very quickly by referring to its definition outlined in formula (1) in section 2.2. Figure A1.1 (a) shows the hotness distribution of lectures.

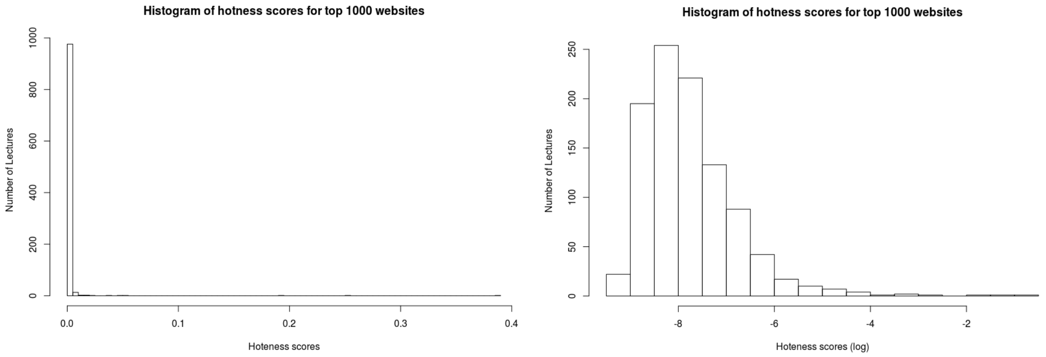


***Figure A1.1. (a)** Histogram of hotness scores for top 1000 lectures* ***(b)** Histogram of log hotness scores for top 1000 lectures*

A log transformation was used to evenly distribute the hotness scores. Figure A1.1 (b) portrays the hotness distribution after a log transformation was applied. Ones, the log transformation was applied, 4 quality classes were generated using quartiles as shown in figure A1.2.

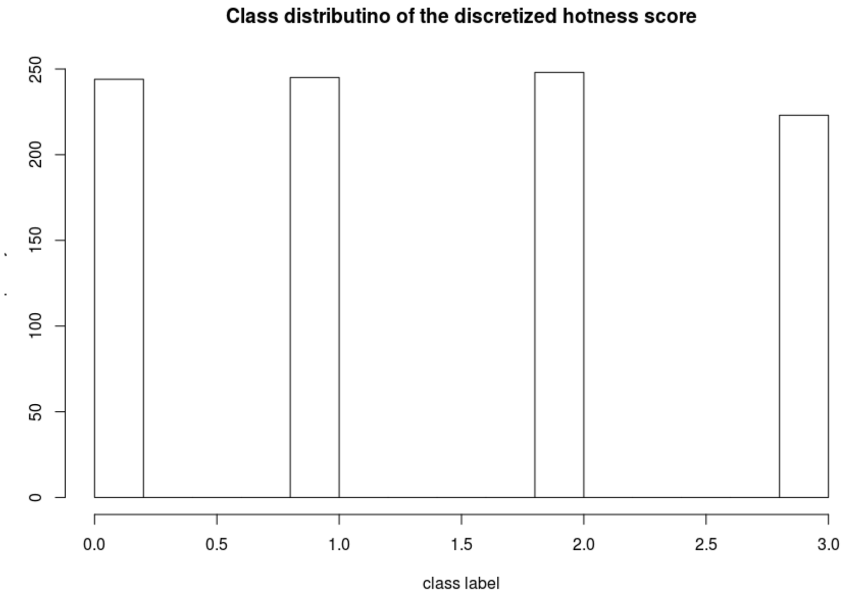


Figure A1.2: *Class distribution after categorising lectures based on log hotness quartiles*

A1.4. Features

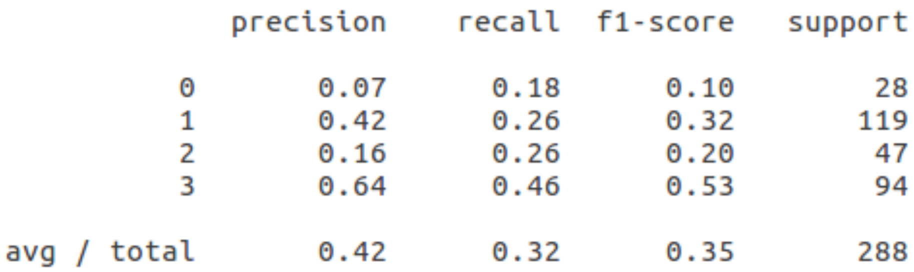
6 features were developed based on the text transcripts and the additional lecture related information we had.

- **Average Sentence length:** mean word count per sentence in text transcript
- **Easiness:** Flesch-Kincaid reading ease test score
- **Duration:** duration of lecture in seconds
- **Fraction complex words:** proportion of complex words in the document
  - Complex word: more than three syllables per word
- **Silence per word:** # of silence tags in transcript per actual word in lecture
- **Title word count:** # of words in the lecture title

A1.5. Results

For each quality class, we trained a binary classifier using logistic regression[3] with a l2-regularization parameter C=1.0. When training the classifier, the dataset was split 70:30 between train and test sets. The regularization parameter was trained by applying 5-fold cross validation.

[3] refer to this model for further information: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html



```
             precision    recall  f1-score   support

          0       0.07      0.18      0.10        28
          1       0.42      0.26      0.32       119
          2       0.16      0.26      0.20        47
          3       0.64      0.46      0.53        94

avg / total       0.42      0.32      0.35       288
```

Figure A1.3: *Classification report of the 4-quality class classifier here class 0: lowest quality... class 3: highest quality*

A1.6. Conclusion

From the classification results, we can see that class 3 (highest) is the only class with marginally positive results. This suggest that the current features are good at detecting high quality material. However, the same doesn’t apply to other classes.

A2. Definitions of Evaluation Metrics

A2.1. Evaluating regression models

In regression model scenario, Metrics such as Root Mean Square Error (RMSE) and Coefficient of determination (R^2) is used to measure the predictive power of models. RMSE and R^2 are geared towards measuring the accuracy of predicting a real value (when y ∊ ℝ).

A2.1.1. Root Mean Square Error (RMSE)

RMSE is a measure that computes the average deviation between true label and predicted value for a given dataset.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (2)$$

As shown in equation (2), RMSE computes the exact error between your datapoints to summarize the total mean error of your model. It is observable from (2) that this metric is scale dependent. If the scale of the data is changed, the error will change.

A2.1.2. Coefficient of Determination (R^2)

Coefficient of determination (referred to as R^2) is another very popular metric for regression models. R^2 represents the proportion of variance of the target variable predictable by the independent variables (features). (3) outlines the definition of R^2.

$$\bar{y} = \sum_{i=1}^{N} y_i$$

$$Var_{tot} = \sum_{i=1}^{N}(y_i - \bar{y})^2$$

$$Var_{reg} = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$$

$$R^2 = \frac{Var_{reg}}{Var_{tot}} \quad (3)$$

This measure range between 0 and 1 where it is close to 0 when very little variance is explained by the model (Hense less predictive power) and close to 1 when most of the variance is explained (leading to high predictive power).

A2.2. Regression for ranking

But in this scenario, the objective is to use a regression model to predict values in such a way that the order of observations is accurate. Therefore, the exact deviation of the prediction from the true value is not important if the global order is preserved.

A2.2.1. Spearman rank coefficient

Ranking metrics are a more suitable metric to measure accuracy in this case. Spearman rank correlation is one of the most used non-parametric metrics when measuring rank correlation between two lists.

$$r_{spearman]} = 1 - \frac{6\sum_{i=1}^{N}(rank(y_i) - rank(\hat{y}_i)))}{N(N^2 - 1)} \tag{4}$$

When the two lists contain distinct integer ranks (*rank(x)*), the correlation coefficient can be computed using the formula in (4) where there are N number of observations in each list.

Spearman Rank Correlation takes a continuous range between –1 and +1 where numbers close to ±1 suggest high rank correlation whereas values close to 0 suggest no rank correlation.

A2.2.2. Kendal-Tau rank correlation coefficient

Kendal-Tau is another non-parametric rank correlation measurement metric that is suitable for measuring how well the ranks are aligned between the true quality of lectures vs. predicted. It is evident from equation (4) that Spearman Correlation requires distinct integer ranks and ties in ranks would introduce complications to (4). Kendal-Tau metric is designed in a way that it is robust to ties.

A2.3. Pairwise Classification Models

Pairwise preference is an approach used to convert a ranking problem into a classification problem. By converting the problem in to a pairwise preference problem, it is possible to express the ranking landscape in more detail using N^2 observations.

There are a few metrics that are popular when evaluating classification models. They are Precision, Recall and Accuracy scores. Before defining these metrics, we define some terminology in Table A1.1:

Actual Label	Predicted Label	Definition	Abbreviation
+1	**+1**	True Positive	TP
-1	**+1**	False Positive	FP
-1	**-1**	True Negative	TN
+1	**-1**	False Negative	FN

Table A1.1: *Cases when comparing actual values with predictions in classification*

A2.3.1. Accuracy Score

Accuracy score quantifies the absolute agreement between the actual labels and the predicted labels in a classification dataset. The definition of Accuracy score is for a dataset of N observations is outlined by equation (5) where the abbreviations are from *Table A1.1*.

$$Accuracy\ Score = \frac{TP + FP}{TP + FP + TN + FN} = \frac{TP + FP}{N} \tag{5}$$

A2.3.2. Precision Score

In the classification context, precision score is the fraction of positively classified observations that are truly positive. The definition is outlined by (6)

$$Precision\ Score = \frac{TP}{TP + FP} \tag{6}$$

A2.3.3. Recall Score

In the classification context, recall score is the fraction of positively classified observations out of all the positive observations in the actual label set. The definition is outlined by (7)

$$Recall\ Score = \frac{TP}{TP + FN} \tag{7}$$