Summer 2018

Matthew Dobbin

Northwestern University

PREDICT 454 Advanced Modelling Techniques

DrivenData: Reboot: Box-Plots for Education

Understanding school budgets using multiclass multilabel classification.

https://www.drivendata.org/competitions/46/box-plots-for-education-reboot/

Contents

1.0 Introduction

Education resource strategies (ERS) is a non-profit organisation that has partnered with more than 40 school systems since 2005 to help them understand and transform the way they use their resources (people, time and money). Their approach integrates data analysis, design and implementation (ERS, 2018).

For ERS to compare budget data across districts, they assign categories to every budget line item. These categories describe what the spending is, what the spending does, which group of students benefit and where the funds come from. This financial coding is a time consuming and labour-intensive task. Creating an algorithm that automates the financial coding paired with human checks will allow ERS to provide valuable insights more quickly and to many more school districts at a lower cost (DrivenData, 2018).

This report details the exploration of the competition data set and the implementation of several multi-class multi-label classification techniques that predict canonical labels to the freeform text budget line items. Tables and figures referenced in the report are contained within the Appendix.

2.0 Literature Review

A review of the literature found that in terms of published peer reviewed journal articles the use of multi-class multi-label classification for budget analysis appears to be a unique application. However, when generalising the search parameters, it was found that there was a vast amount of literature regarding multi-class classification as well as using natural language processing techniques to help classify freeform text documents.

In the article "Applying machine learning in accounting research" the authors discuss that accounting research is often limited due to the manual labour required to analyse each artefact. Their research found that machine learning techniques that use natural language processing can "reliably and efficiently classify large numbers of unstructured documents" (Bogaerd, M. V., & Aerts, W. 2011).

Natural language processing can be used in a wide variety of applications. In the medical field, research published by Szlosek and Ferretti investigated the performance of machine learning techniques for automating the evaluation of clinical decisions in electronic medical record systems. The machine learning program used natural language processing techniques including vectorization and tokenization to pre-process and analyse freeform text documents such as clinical reports and physician notes. The authors evaluated the performance of a k-nearest neighbors classifier, a decision tree classifier and a c-support vector classifier. They found the k-nearest neighbors classifier to be the least accurate (Szlosek, D. A., & Ferretti, J. M. 2016).

The journal article "Systematic examination of the incorporation of class relationships via multilabel, multiclass, hierarchical classification" discusses several relevant points for this problem. Firstly, poor performance can occur in multi-class classification problems due to transforming the problem into a series of one class problems. This is done by determining each class without respect to the results from the other classes. Secondly, accuracy is poorly suited to being used as the single performance metric for the classifier and can lead to models that assign no labels. Finally, in their investigation into multi-label classification performance, the authors found that for their application the models with the best performance incorporated the random forest classifier (Daisey, K., & Brown, S. D. 2017). The performance of a random forest classifier for this completion is discussed in Section 4.5 of this report.

3.0 Competition and Data Exploration

The training dataset which contains the features and the labels consists of 400277 line items (rows). The test data set contains just the features and consisted of 50064 line items. Each row in the budget is mainly comprised of freeform text columns. There are two float columns. One is related to the number of full time workers. The other is the total cost of expenditure. A full list of the features in the data set and their descriptions is provided in Table 1.

For each line item, ERS attaches one label from each of the nine different categories. These labels and categories are shown in Table 2. The number of labels in each category has been visualised and is shown in Figure 1. The category 'function' has the most labels.

The goal of the competition is to predict a probability for each possible label for each line item in the test set. To do this, the column structure was flattened so that for each label in each category had its own column with a probability assigned. Due to this flattened structure, the submission file consisted of 50064 rows by 104 columns.

3.1 Learning Python

This project was the author's first attempt at using Python for predictive modelling. This competition was chosen as it was an opportunity to explore natural language processing. To assist with the learning process, several tutorials from Datacamp were completed. These included Supervised Learning with Scikit-learn, Machine Learning with Experts: School Budgets and Deep Learning in Python (Datacamp, 2018).

Codes sections from the tutorial for building a logistic regression pipeline that merges numerical and text features for the competition have been copied. Modifications have been made in order to optimize performance, explore and visualise the data set and the implementation of other classification model types.

3.2 Pre-Processing

The pre-processing of the data was conducted by creating two separate pipelines. There was one pipeline for the two numerical features and the second pipeline for processing of the text features. The summary statistics for the numerical variables are shown in Table 3 and boxplots for the variables are shown in Figure 2 and Figure 3.

Approximately 70% of the values for the FTE variable are missing. From the boxplot it appears as though there is a large number of unexpected values as well. The expected values for FTE would be between 0-1 as it is a percentage of full time work. Values outside of the range of 0-1.2 were changed to NaN and then the missing values were imputed using the mean value.

Less than 5% of the values were missing for the total cost of expenditure variable. The maximum value of 129,700,000 could potentially be an outlier. Without having a full understanding of the accounting system and expected values for the school budget, it is hard to determine if it is an unexpected value. An assumption was made that a reasonable maximum amount for a budget line item would be 1,000,000. Missing values in the numerical data were imputed using the mean value.

3.3 Text Data Exploration

Columns in the data set that were not labels or numerical variables were combined into a single vector for each row. A function was used to remove stop words (commonly used words such as "the") and then tokenize the text into unigrams, bigrams and trigrams. To obtain an insight into the text data, a frequency distribution of the top fifty tokens was created. This plot is shown in Figure 4.

The top five most frequent tokens were 'general', 'regular', 'fund', 'teacher' and 'services'. The most frequent bigram was 'employee benefits' and this was the 14th most frequent token overall. A word cloud was used to visualise the text data in a different fashion. This word cloud is shown in Figure 5.

4.0 Classification Modelling

The following section describes the different multi-class multi-label classification techniques that were used to predict canonical labels for the freeform text budget line items. The techniques include one versus rest logistic regression, decision tree and random forest.

4.1 Scoring Metric

Multi multi class log loss was chosen as the metric to evaluate the classifier performance for this competition. As discussed in the literature review, accuracy is not always a good indicator of overall classifier performance. It measures the number of predictions where the predicted value equals the actual value. Log loss measures error with the aim of minimising how much the prediction varies from the actual value. The formula to calculate this metric is as follows:

$$Multi\ multi \log loss = \frac{1}{\mathrm{K}}\sum_{k=1}^{K}[-\frac{1}{N}\sum_{n=0}^{N}\sum_{c=1}^{C} y_{k,c,n}\log(\hat{y}_{k,c,n})]$$

- K is the number of dependent variables
- N is the number of rows being evaluated
- C is the number of class values k can take on

Unfortunately, currently no metrics in sklearn.metric support the multi-label multi-class classification output (Scikit-learn, 2011). A function to return the log loss values for the validation test was created however there are doubts about the accuracy of this function.

4.2 Modelling Pipeline

Two pipelines were created to pre-process the text and numerical variables separately. For the text pipeline the combined text vector was selected and then the CountVectorizer() function was used to tokenize the strings. The number of text features selected for the model was varied by selecting the 'k' best features based on the chi-squared statistical test. The two pipelines were joined using the FeatureUnion function and then the features were scaled.

4.3 Logistic Regression

A logistic regression model can be used in multi-class multi-label classification when the one versus all approach is used. This approach fits a separate classifier for each of the dependant variables. Multiple logistic regression models were fitted. The parameters used and the performance statistics for these models is shown in Table 4.

The parameters that were varied for the different models were the select the 'k' best features and the types of n-grams that were used in the model. The computational time for the models ranged between 5-11 minutes. The model that only used 50 text features was the worst performing.

Doubts about the accuracy of the log loss function for the validation test set were identified. The k=50 model has a smaller log loss value than the k=1000 model. When submitting the test set predictions for the competition, the k=1000 model outperformed k=50 as well as having a significant difference between the validation set and test set log loss scores (4.88 vs. 0.64 respectively).

The plot in Figure 6 shows the logistic regression test set log loss values versus the number of text features selected in the model. From the plot it can be seen that adding additional features after 1000 text features did not improve the classifier performance. One surprising outcome when reviewing the models' performances was that the model with only unigram tokens outperformed the model that had unigram and bigrams. The features and the corresponding coefficients for the k=1000 and unigram only model were extracted from the pipeline and exported. The file titled 'Logisitic_Coefficients.csv' was part of the final submission.

4.4 Decision Tree

A decision tree classifier was the second type of model that was fitted. It was decided to use k=1000 and unigram only tokens based on the results from the logistic regression models. The decision tree log loss value of 10.2 was quite high. When reviewing the predictions on the test set, it was found that the probabilities outputted were only 0 and 1 despite using the pipe.predict_proba() function. This would also explain why the validation test set scored a high accuracy value of 0.91.

4.5 Random Forest

Random forest is an ensemble learning method that constructs a multitude of decision trees and then outputs the class that is the mode of the classes. The random forest method corrects the habit of decisions trees overfitting the training data. Two random forest models were fitted. The first model used the default Sci-kit learn parameters. The computational time for the first random forest model was approximately three hours.

The second random forest model that was implemented used a grid search approach and k-fold cross validation. Only a few parameters were selected to be tuned due to the computational requirements. The parameters for the best estimator model are shown in Table 6. The default option in Sci-kit learn for the number of trees in the forest is ten (Scikit-learn, 2011). The grid search found that the best estimator had 100 trees in the forest. With the hyperparameter tuning the log loss score improved from 1.69 to 0.98.

5.0 Results Comparison

It was decided to use the test score submission value to compare the performance of the different classifiers. This was due to identifying large discrepancies between the validation set and the test set submissions log loss values. Table 7 shows a comparison of the overall performance for the different classifier types.

The logistic regression model was the best performing model in terms log loss score. The computational requirements for the random forest grid search was quite high at 9hrs and this was for a grid search that was not that extensive.

6.0 Conclusion

In terms of overall competition performance, the logistic regression model that used the best 1000 unigram text features resulted in the best test submission score. The model’s test set log loss value of 0.6359 is currently placed ninth (was seventh) out of 756 competitors. A screenshot of this result is shown in Figure 7. The lowest log loss score is currently sitting at 0.36.

References

Bogaerd, M. V., & Aerts, W. (2011). Applying machine learning in accounting research. *Expert Systems with Applications, 38*(10), 13414-13424.

Daisey, K., & Brown, S. D. (2017). Systematic examination of the incorporation of class relationships via multilabel, multiclass, hierarchical classification. *Journal of Chemometrics, 31*(5).

DataCamp. (2018). Data Scientist with Python Track | DataCamp. Retrieved August 02, 2018, from https://www.datacamp.com/tracks/data-scientist-with-python

DrivenData. (2018). Reboot: Box-Plots for Education. Retrieved August 01, 2018, from https://www.drivendata.org/competitions/46/box-plots-for-education-reboot/

ERS. (2018). Education Resource Strategies: Urban School Resource Organization and Transformation. Retrieved August 14, 2018, from https://www.erstrategies.org/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R*.

Scikit-learn. (2011). 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier¶. Retrieved September 01, 2018, from http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Scikit-learn. (2011). 1.12. Multiclass and multilabel algorithms¶. Retrieved September 01, 2018, from http://scikit-learn.org/stable/modules/multiclass.html

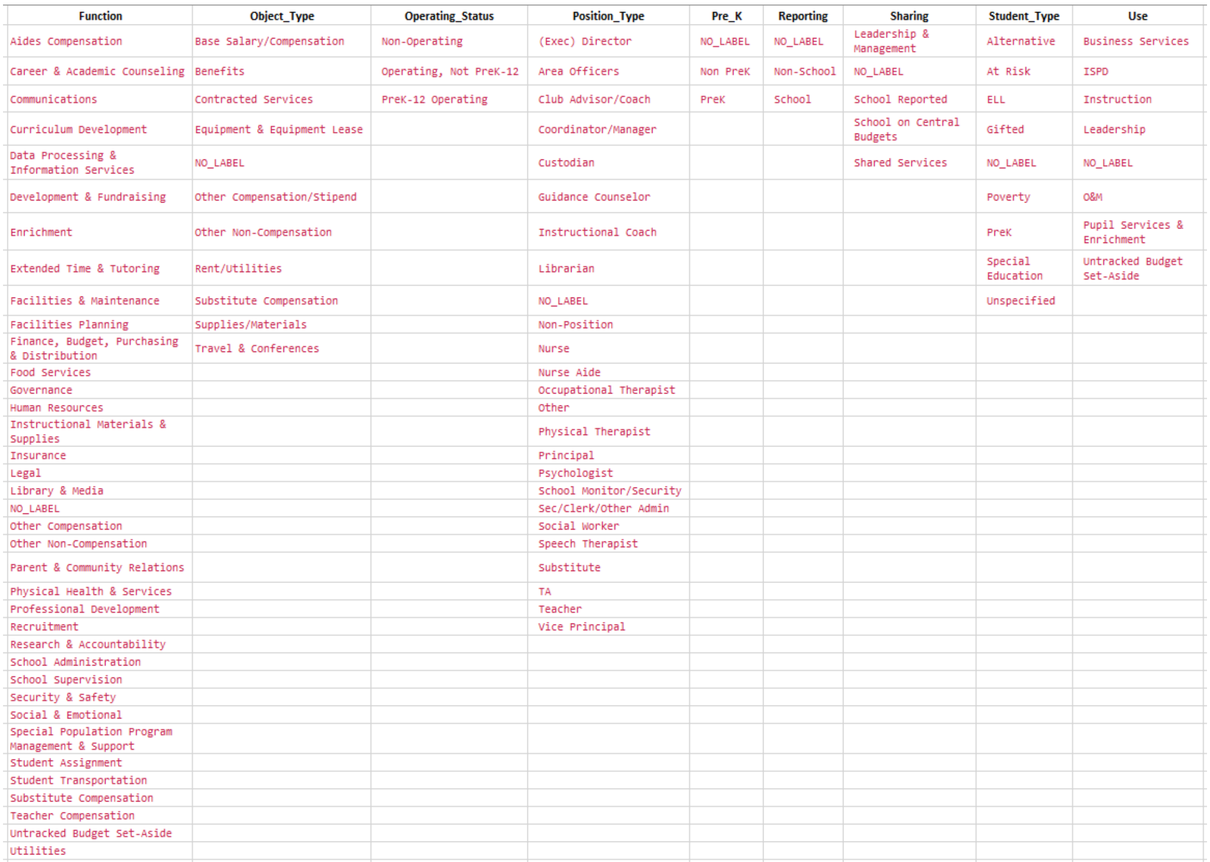
Szlosek, D. A., & Ferretti, J. M. (2016). Using Machine Learning and Natural Language Processing Algorithms to Automate the Evaluation of Clinical Decision Support in Electronic Medical Record Systems. *EGEMs (Generating Evidence & Methods to Improve Patient Outcomes), 4*(3), 5.

Appendix A

Table 1 - Descriptions of the features contained within the budget.

FTE float - If an employee, the percentage of full-time that the employee works.
Facility_or_Department - If expenditure is tied to a department/facility, that department/facility.
Function_Description - A description of the function the expenditure was serving.
Fund_Description - A description of the source of the funds.
Job_Title_Description - If this is an employee, a description of that employee's job title.
Location_Description - A description of where the funds were spent.
Object_Description - A description of what the funds were used for.
Position_Extra - Any extra information about the position that we have.
Program_Description - A description of the program that the funds were used for.
SubFund_Description - More detail on Fund_Description
Sub_Object_Description - More detail on Object_Description
Text_1 - Any additional text supplied by the district.
Text_2 - Any additional text supplied by the district.
Text_3 - Any additional text supplied by the district.
Text_4 - Any additional text supplied by the district.
Total float - The total cost of the expenditure.

Table 2 - List of the 9 ERS categories and the associated labels.



Function	Object_Type	Operating_Status	Position_Type	Pre_K	Reporting	Sharing	Student_Type	Use
Aides Compensation	Base Salary/Compensation	Non-Operating	(Exec) Director	NO_LABEL	NO_LABEL	Leadership & Management	Alternative	Business Services
Career & Academic Counseling	Benefits	Operating, Not PreK-12	Area Officers	Non PreK	Non-School	NO_LABEL	At Risk	ISPD
Communications	Contracted Services	PreK-12 Operating	Club Advisor/Coach	PreK	School	School Reported	ELL	Instruction
Curriculum Development	Equipment & Equipment Lease		Coordinator/Manager			School on Central Budgets	Gifted	Leadership
Data Processing & Information Services	NO_LABEL		Custodian			Shared Services	NO_LABEL	NO_LABEL
Development & Fundraising	Other Compensation/Stipend		Guidance Counselor				Poverty	O&M
Enrichment	Other Non-Compensation		Instructional Coach				PreK	Pupil Services & Enrichment
Extended Time & Tutoring	Rent/Utilities		Librarian				Special Education	Untracked Budget Set-Aside
Facilities & Maintenance	Substitute Compensation		NO_LABEL				Unspecified	
Facilities Planning	Supplies/Materials		Non-Position					
Finance, Budget, Purchasing & Distribution	Travel & Conferences		Nurse					
Food Services			Nurse Aide					
Governance			Occupational Therapist					
Human Resources			Other					
Instructional Materials & Supplies			Physical Therapist					
Insurance			Principal					
Legal			Psychologist					
Library & Media			School Monitor/Security					
NO_LABEL			Sec/Clerk/Other Admin					
Other Compensation			Social Worker					
Other Non-Compensation			Speech Therapist					
Parent & Community Relations			Substitute					
Physical Health & Services			TA					
Professional Development			Teacher					
Recruitment			Vice Principal					
Research & Accountability								
School Administration								
School Supervision								
Security & Safety								
Social & Emotional								
Special Population Program Management & Support								
Student Assignment								
Student Transportation								
Substitute Compensation								
Teacher Compensation								
Untracked Budget Set-Aside								
Utilities								

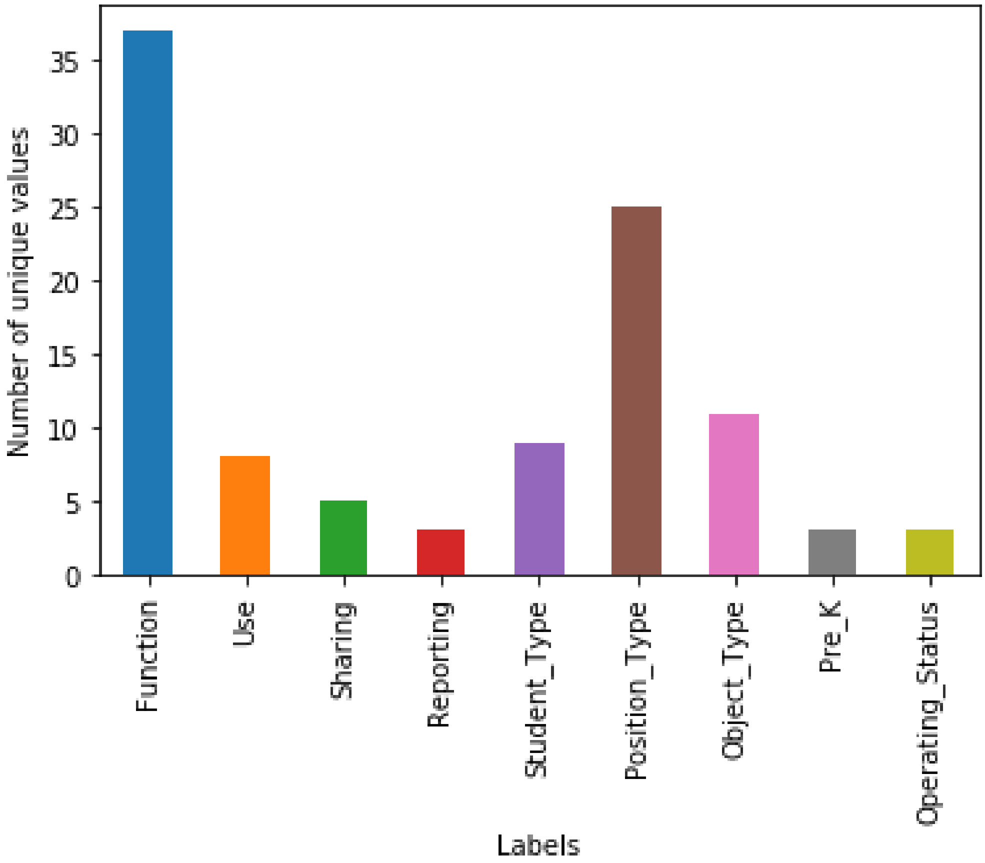


Figure 1 - Count of labels for each ERS category.

Table 3 - Numerical variables summary statistics.

	FTE	Total
count	126071	395722.00
mean	0.426794	13105.86
std	0.573576	368225.40
min	-0.087551	-87,466,310.00
25%	0.000792	73.80
50%	0.130927	461.23
75%	1	3652.66
max	46.8	129,700,000.00

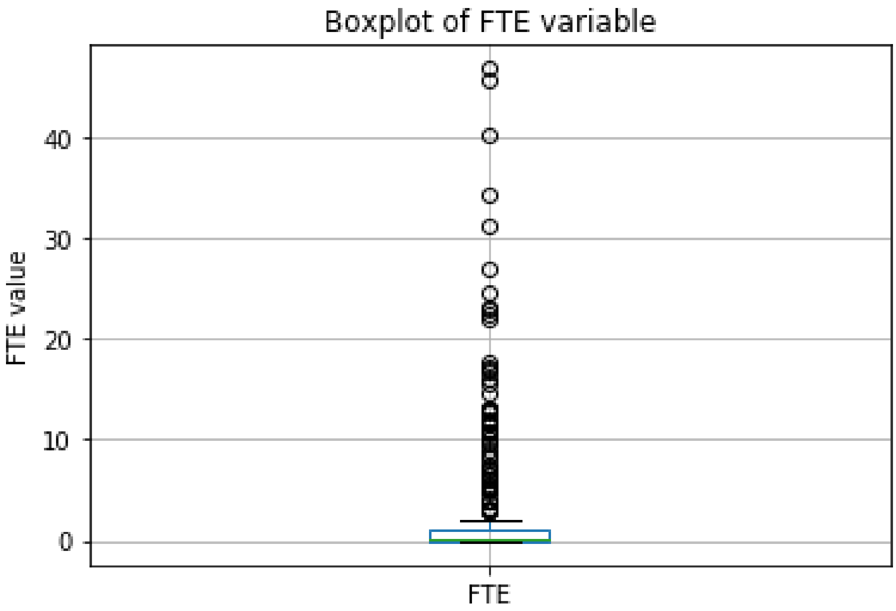


Figure 2 - Boxplot of FTE numerical variable.

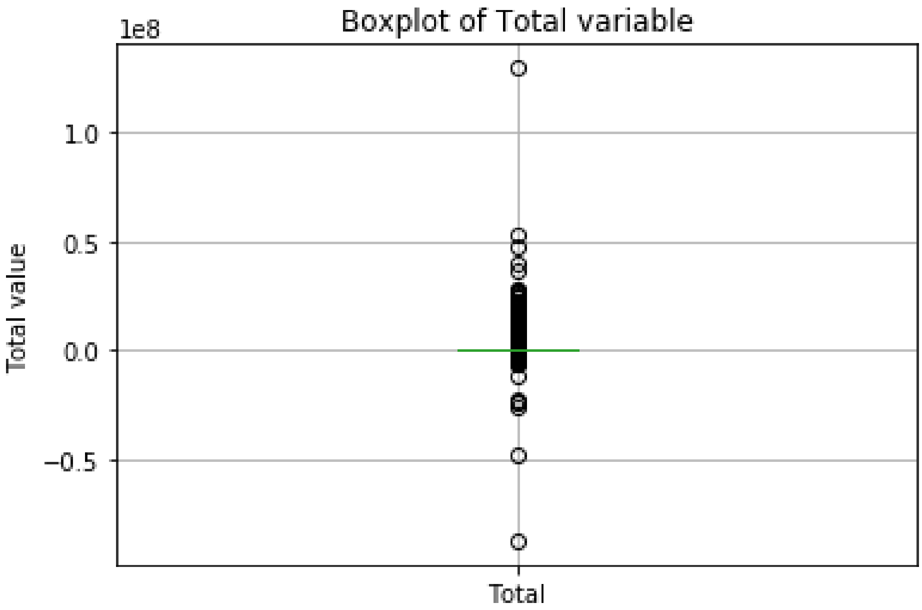


Figure 3 - Boxplot of Total expenditure numerical variable.

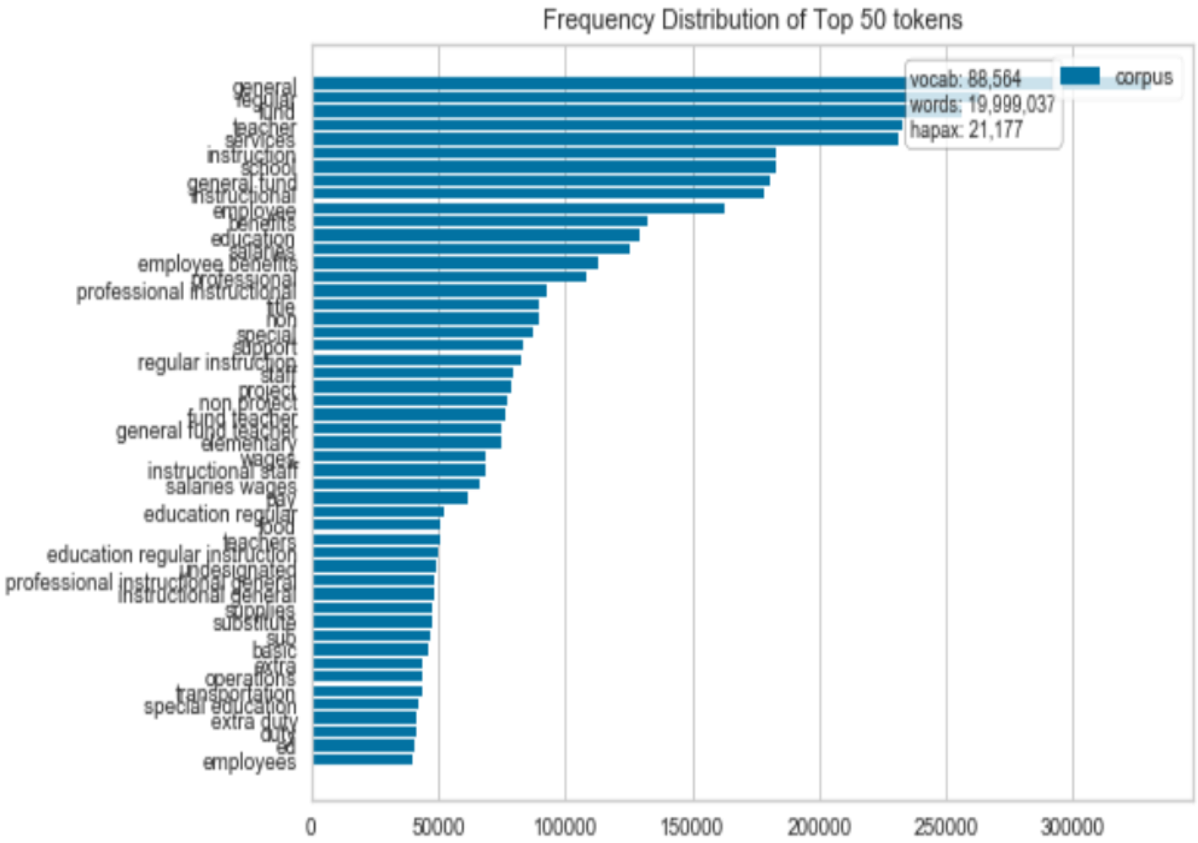


Figure 4 - Frequency Distribution of top 50 tokens.

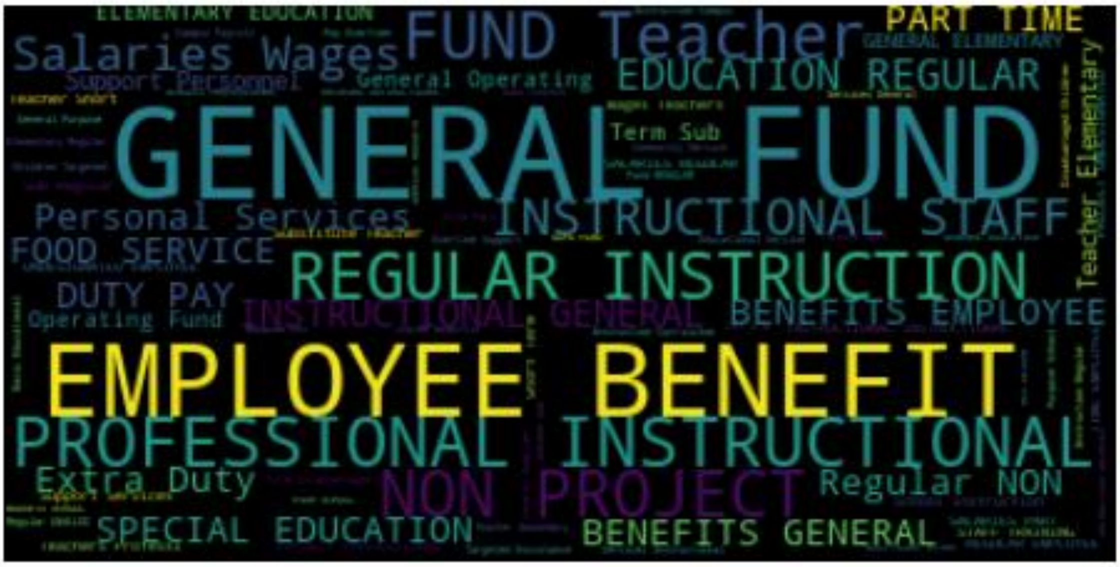


Figure 5 - Word cloud of non-label text columns.

Table 4 - Performance statistics for the multiple Logistic Regression models that were fitted.

Model	Features (chi_k)	Computation Time	Validation Accuracy	Validation Log Loss	Log Loss
Logistic	50	5min	0.327	3.64	0.9212
Logistic	300	9min	0.65	4.697	0.7558
Logistic	1000	8min	0.792	4.879	0.6359
Logistic	2000	10min	0.817	4.896	0.6407
Logistic	3000	11min	0.819	4.897	0.6387

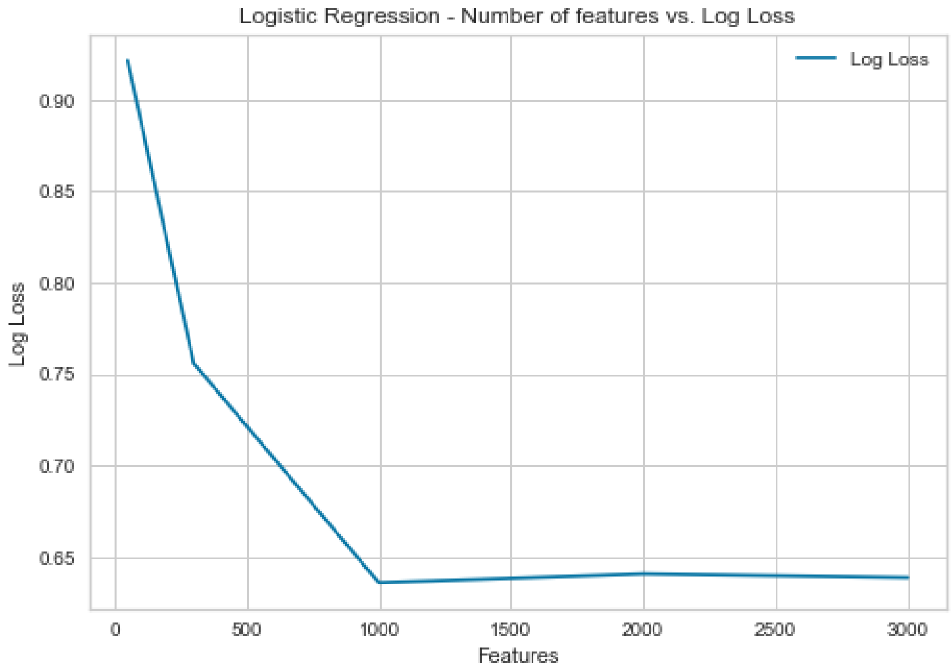


Figure 6 - Plot of Logistic Regression Test Set Log Loss versus the number of text features selected in the model.

Table 5 - Comparison of the logistic regression model when the number of n-grams were varied.

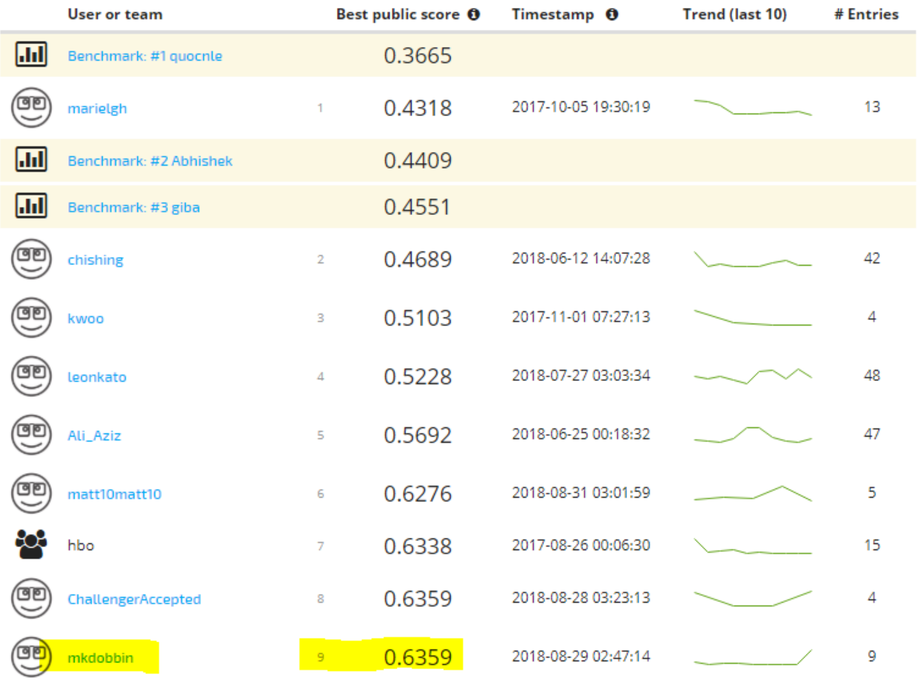
Model	Features (chi_k)	n_gram	Log Loss
Logistic	1000	ngram_range = (1,1)	0.6359
Logistic	1000	ngram_range = (1,2)	0.7329
Logistic	1000	ngram_range = (1,3)	0.8979

Table 6 - Grid search parameters that resulted in the best random forest model.

Tuning Parameter	Best Estimator Value
max_depth	10
min_samples_leaf	5
n_estimators	100

Table 7 - Comparison of overall performance for the different classifier types.

Best Models	Computational Time	Log Loss
Logistic	11min	0.64
Random Forest with Grid Search	9hrs 40min	0.98
Random Forest	3hrs	1.69
Decision Tree	3min	10.2



User or team		Best public score	Timestamp	Trend (last 10)	# Entries
Benchmark: #1 quocnle		0.3665			
marielgh	1	0.4318	2017-10-05 19:30:19		13
Benchmark: #2 Abhishek		0.4409			
Benchmark: #3 giba		0.4551			
chishing	2	0.4689	2018-06-12 14:07:28		42
kwoo	3	0.5103	2017-11-01 07:27:13		4
leonkato	4	0.5228	2018-07-27 03:03:34		48
Ali_Aziz	5	0.5692	2018-06-25 00:18:32		47
matt10matt10	6	0.6276	2018-08-31 03:01:59		5
hbo	7	0.6338	2017-08-26 00:06:30		15
ChallengerAccepted	8	0.6359	2018-08-28 03:23:13		4
mkdobbin	9	0.6359	2018-08-29 02:47:14		9

Figure 7 - Screenshot of the best submission from the competition website.