APPLIED DATA SCIENCE CAPSTONE: PREDICTING AND MODELLING THE SCHOOL PERFOMANCE FOR 2020

Sabelo Yalezo

In fulfilment of IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

2020 August 03

Table of Contents

INTRODUCTION

Background

Education is a basic right of every Human on this Planet, and the governments of all countries ensure to spread Education. It is a weapon to improve one's life and certainly determines the quality of an individual's life. It seems like many poor people improve their lives with the help of Education. As it teaches the value of discipline to individuals and enables individuals to express their views clear manner. Educated people are quite likely to convince people to their point of view, off which is desirable trait from employment and entrepreneurship. Furthermore, it helps in spreading and transferring knowledge from one generation to another in our societies.

Significance and rational

While education improves one's knowledge, skills and develops the individual personality and attitude, there is a lack of quality Teachers amongst the school in South Africa. The main contributor to this, relies on how the education department attracts quality Educators. With lot of good quality Teachers opting for Urban areas as opposed to rural areas. Then question arise: If most of our quality Teachers are in Urban schools, does this mean the children in poor areas do not deserve quality education?

Research objectives are as follows: to determine the effects of Quintile on school performance.

1. To predict the school performance based on historic school performance data.
2. To determine school that need urgent intervention before end of the year exams.
3. By finding 3 we would have found the best performing schools. And create 3 clusters on a MAP.

Interested Stakeholders

Hence in this project we recommend to department of education the top 10 schools per district that need urgent intervention based on the quality of education pupil received. This is so that they can place and employ teachers based on the needs of the schools. So that quality of education is the same across. This is done by exposing the current school rating and the areas of interest around the school town.

DATA WRANGLING

Based on definition of our problem, requirements that will influence our decision are: first is the number of existing schools in the neighbourhood, whether it's a private or public school. Secondly, number of schools in the neighbourhood, using average income (Quintile) if any. Lastly, distance of schools from city centre. We decided to use the capital city of EC province locations, to visualize and plot our neighbourhoods. For each district we will check the best school's recommender for teacher to pursue career on, based on the school rating and areas of interest around the district. We mainly focus on the Eastern Cape province, where there is low pass rate.

While education improves one's knowledge, skills and develops the individual personality and attitude, there is a lack of quality Teachers amongst the school in South Africa. The main contributor to this, relies on how the education department attracts quality Educators. With lot of good quality Teachers opting for Urban areas as opposed to rural areas. Then question arise: If most of our quality Teachers are in Urban schools, does this mean the children in rural areas do not deserve quality education? Hence in this project we recommend to department of education the top 10 schools per district that need urgent intervention based on the quality of education pupil received. This is so that they can place and employ teachers based on the needs of the schools. So that quality of education is the same across. This is done by exposing the current school rating and the areas of interest around the school town.

Data Collection

The capital City of the EC province addresses is extracted from Google Maps API reverse geocoding**.** The rest of the data contains school performance info, district names, latitude and longitude, suburb and other unrelated info., as details below:

- ***School location Data***: To get the neighbouring information about the school. This include the nearest towns, municipality around. Furthermore, whether the school is public or private, can be found http://ecdoe.co.za/files/resources/resource_260.xlsx.

Sector	Phase_DoE	STATUS	Institution_Type	Township_Village	...	DMunName	LMunName	Telephone	Facsimile	GIS_Long	GIS_Lat	Section21
Public	PRIMARY SCHOOL	OPERATIONAL	ORDINARY SCHOOL	NaN	...	Cacadu	Sundays River Valley	422330419.0	422330419	25.71395	-33.52716	Yes
Public	COMBINED SCHOOL	OPERATIONAL	ORDINARY SCHOOL	NaN	...	Nelson Mandela Bay	Nelson Mandela Bay	414669018.0	414669018	25.58508	-33.86708	No
Public	PRIMARY SCHOOL	OPERATIONAL	ORDINARY SCHOOL	NaN	...	Cacadu	Camdeboo	498460031.0	498460045	24.06113	-32.47425	Yes
Public	SECONDARY SCHOOL	OPERATIONAL	ORDINARY SCHOOL	NaN	...	Cacadu	Camdeboo	498460353.0	498460353	24.06183	-32.47385	Yes
Public	PRIMARY SCHOOL	OPERATIONAL	ORDINARY SCHOOL	NaN	...	Nelson Mandela Bay	Nelson Mandela Bay	414533461.0	414533461	25.56570	-33.93000	No

- ***School Performance Data***: contains the historic performance data from 2017. It outlines the number of Quintile, students wrote, achieve and failed, together with the school pass rate. https://www.sace.org.za/assets/documents/uploads/sace_54588-2020-01-10-2019%20NSC%20School%20Performance%20Report.pdf

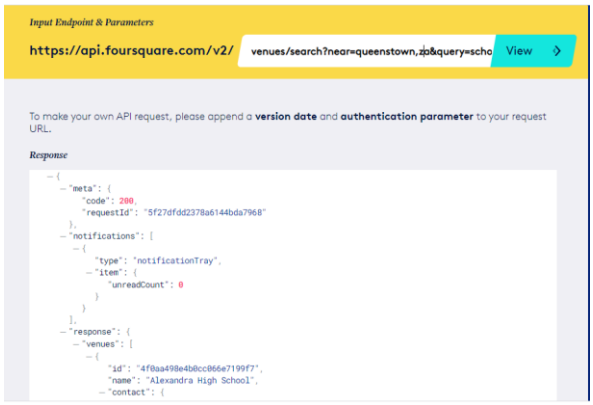
	District Name	EMIS No	Centre No	Centre Name	Quintile	2017_No Progressed	2017_Wrote	2017_Achieved	Percentage_2017_Achieved	2018_No Progressed	2018
3		200500039	4241001	A D TSHAYINGCASECONDARY SCHOOL	3	12	252	194	77.0	55	
4		200500013	4241002	BALENI SENIOR SECONDARYSCHOOL	1	46	101	60	59.4	16	
5		200500041	4241003	BIZANA SENIOR SECONDARYSCHOOL	3	88	355	265	74.6	91	
6		200501456	4241004	CANGCI COMPREHENSIVETECHNICAL HIGH	1	46	87	58	66.7	37	
7		200501404	4241005	CHIEF DUMILE SENIORSECONDARY SCHOOL	1	52	126	115	91.3	71	

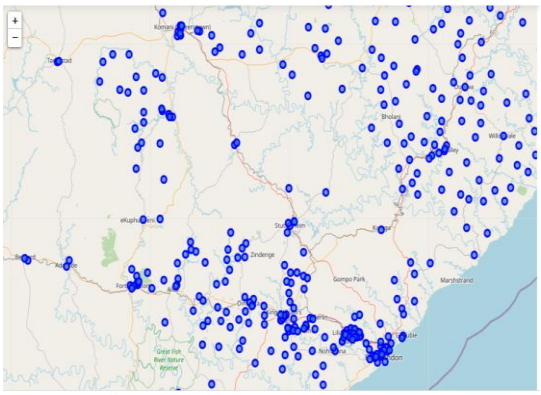
For the above data souced we used Camelot to read the pdf tables and pandas to store data. The data was manipulated manipulate and clean, and some of the features were deleted.

Methodology

In this section we detail on how the project was designed and executed. We used different data sources to gather information as outlined in the previous section. After the processing of data, we determined the important features for modelling suing both classifying and Linear modelling algorithms. We used Linear regression, Ridge, Rando Forest and Boost as the linear models. While Support Vector Machine, Decision Tree, K-Nearest Neighbour, Logistic regression were used as classifiers. The approach to use both models is motivated by the fact that we wanted to identify the school that improved performance and be able to quantify the magnitude of the improvement whether positive or negative. Lastly, we used the K-means algorithm to segmentation schools using the historical data. By feature segmentation we partition a school into groups that have similar characteristics. These partition results in 3 cluster : best school, worst performing school and average performing school. We need to use a classification algorithm. Infect we need to use cluster. After the wrangling of data , EDM

and modelling, we use Foursquare REST APIS services and Folium to analyse and create clusters. Based on the map created teachers to create 3 clusters on the Map. Afterall the project will trigger attraction to all different parts of province, as such attract and retain good quality teacher. It will show them closest Towns or Cities they can reside on while teaching in different districts and town.





To achieve the project objectives and drawn solution to the problems identified, we used a hybrid of the *Sample*, *Explore*, *Modify*, *Model*, *and Assess (SEMMA)* and CRISP data mining approach.

EXPLORATORY DATA ANALYSIS

To prepare the data for modelling to do EDM we have used different inferential statistic and hot encoding and data normalization. Ranging from correlation coefficient, mean and median.

Calculation of target variable

We create *Status* based on the 2018 and 2017 school performance data. The *Status=1*. If the improvement in the historic data is greater or equal to zero , else the *Status=0,* if the average improvement is less than zero. This done by comparing 2019 results and the historic results. The Status features is used as the target variable for supervised classification model outline in the methodology. To quantifying the magnitude of the improvement using linear models, we set the target variable as '*Percentage_achieved_2019*".

Determining the outliers

The data using School pass rate vs Quantile for the year 2017, 2018, and 2019. The relation between the school-performance and quintile, is one of the objectives we wanted to resolve in this project.

The hypothesis states that the schools that are from higher Quintiles where the neighbourhood income is higher, tend to perform better than the schools in the marginalised neighbourhood

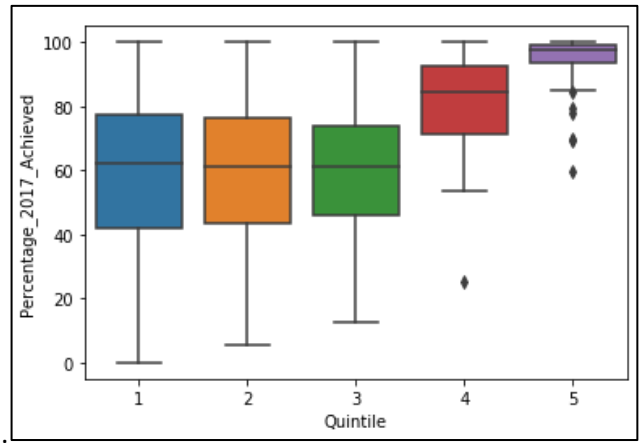


Figure 1. School Pass Rate vs Quintile, 2017

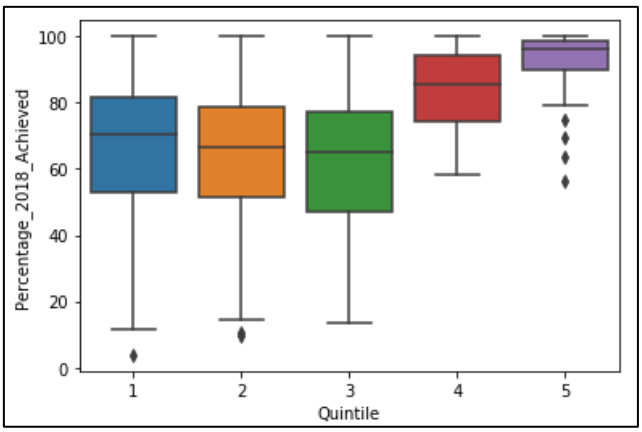


Figure 2.. School Pass Rate vs Quintile, 2018

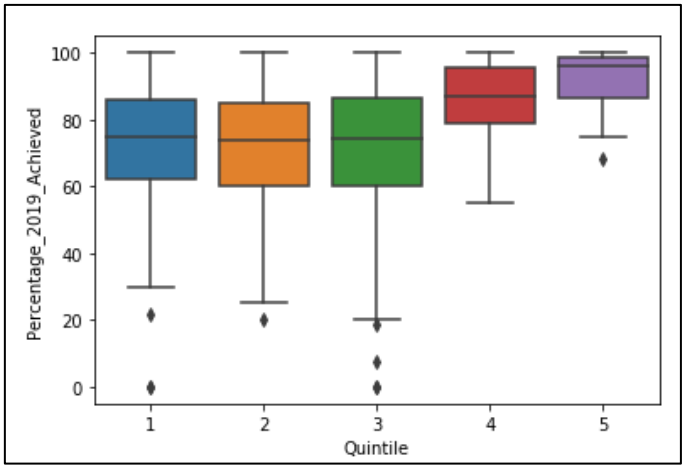


Figure 3.. School Pass Rate vs Quintile, 2019

From the above chart we see that the Quintile (4&5) have an average median of 85, 95 % respectively. This proves our hypothesis that the schools in upper Quintile perform better, even though there are outliers. While Quintile (1,2,3) have median of 78% which has seen improvement looking from 2017.

Determining the relation between average pass rate and average improvement

For the upper Quintile the improvement is close to zero because they are already giving a mean more than 95 %, hence the means is centred around zero. We see the lower Quintile have an average improvement of 9%, which is a median. There fore we can conclude that initiatives are being taken in those schools to improve pass rate.

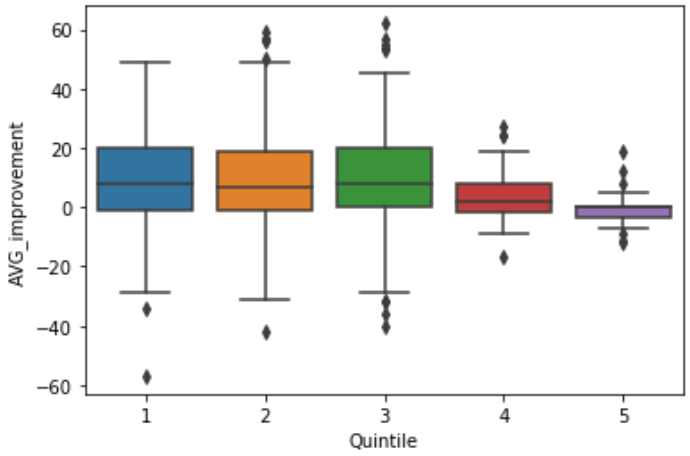


Figure 4. Percentage Improvement between 2018 vs 2017

While according to figure 4 above, the lowe quntile schools have an improved pass rate, thera are some schools that are getting worse, as shown in the digram below:

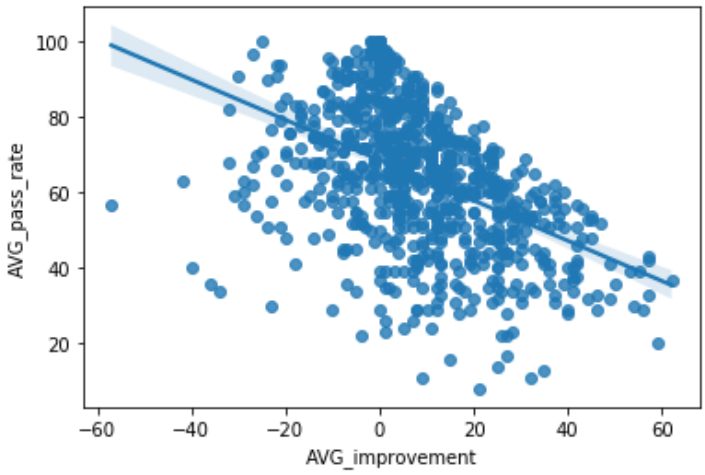


Figure 5.DIstribution of Pass Rate improvement per Quintile

Figure 5, Shows the distibution of pass rate vs improvement. The scatter plot above visualize the extend to which some school aahveing detrpriated perfromance results ven thoough they are on upper quintile.

Feature selection

As outlined in the methodology that we use a hybrid SEMMA and CRISP, the following feature we selected for the modelling using the correlation matrix and heat map. We wanted to predict the pass-rate, we check for feature that related with "*percentage_2018_achieved*".

Table 1. Correlation Matrix

	improvement_2018	Quintile_5	Quintile_4	Quintile_3	Quintile_2	Quintile_1	2019_Wrote	2018_Wrote	2017_Wrote	2017_Achieved	F
improvement_2018	1.000000	-0.079660	-0.032104	-0.064800	-0.016319	0.146533	-0.123152	-0.155873	0.022496	-0.173872	
Quintile_5	-0.079660	1.000000	-0.050960	-0.196556	-0.129633	-0.133388	0.177027	0.167884	0.141793	0.298497	
Quintile_4	-0.032104	-0.050960	1.000000	-0.185690	-0.122467	-0.126014	0.042515	0.024749	-0.004797	0.051755	
Quintile_3	-0.064800	-0.196556	-0.185690	1.000000	-0.472360	-0.486044	-0.096814	-0.077136	-0.058749	-0.086285	
Quintile_2	-0.016319	-0.129633	-0.122467	-0.472360	1.000000	-0.320557	-0.057174	-0.055520	-0.071629	-0.087488	
Quintile_1	0.146533	-0.133388	-0.126014	-0.486044	-0.320557	1.000000	0.055946	0.045125	0.067638	0.007335	
2019_Wrote	-0.123152	0.177027	0.042515	-0.096814	-0.057174	0.055946	1.000000	0.875753	0.762190	0.843114	
2018_Wrote	-0.155873	0.167884	0.024749	-0.077136	-0.055520	0.045125	0.875753	1.000000	0.793336	0.845316	
2017_Wrote	0.022496	0.141793	-0.004797	-0.058749	-0.071629	0.067638	0.762190	0.793336	1.000000	0.877139	
2017_Achieved	-0.173872	0.298497	0.051755	-0.086285	-0.087488	0.007335	0.843114	0.845316	0.877139	1.000000	
Percentage_2017_Achieved	-0.571743	0.327731	0.176089	-0.090104	-0.067639	-0.083047	0.337468	0.283173	0.118586	0.455178	
2018_Achieved	-0.115448	0.272412	0.063562	-0.110968	-0.070038	0.025911	0.892494	0.945063	0.771816	0.900871	
Percentage_2018_Achieved	0.129597	0.294859	0.186662	-0.157511	-0.083556	0.021251	0.338005	0.234880	0.227441	0.419525	
Status	0.514648	-0.025235	0.029143	-0.044469	-0.013347	0.062679	-0.040356	-0.050901	0.092227	-0.057152	
2019_Achieved	-0.118468	0.254140	0.062920	-0.114532	-0.068081	0.037686	0.968968	0.867695	0.760288	0.882856	
Percentage_2019_Achieved	-0.033103	0.231199	0.144662	-0.071654	-0.082847	-0.024638	0.208568	0.226466	0.211048	0.340771	
AVG_pass_rate	-0.264279	0.345949	0.200527	-0.135979	-0.082508	-0.037157	0.373311	0.287450	0.188144	0.484509	
AVG_improvement	0.999988	-0.079604	-0.031941	-0.064919	-0.016386	0.146626	-0.123328	-0.155822	0.022627	-0.173922	
Distance_Closest_Town	0.128706	-0.185168	-0.183142	-0.373547	0.157874	0.454209	-0.067501	-0.084277	-0.054301	-0.129973	

Hence the following feature were selected for the model: *Percentage_2017_Achieved, AVG_pass_rate,Percentage_2018_Achieved,Quintile_5,improvement_2018,AVG_improveme*

nt, ***Status and 2019_Wrote .*** We chose these columns because they had high importance in predicting the magnitude of pass rate improvement.

MODELLING and EVALUATION

Modelling the school performance improvement classifier

We build our model using Logistic Regression , Ridge, Random Forest and XGBoost from Scikit-learn package. The Ridge was as a regularization technique for Linear regression to solve the overfitting problem in machine learning models. For both Random forest and XGboost we used grid search CV and Early stopping, respectively to optimise the models. The performance of linear model was 57%. This means that we could only estimate 57% of test data accurately. This was not optimum accuracy we could get from the models, but it is due to data used. Some of the of determining factors for school-performance were not found, hence the optimization and regulation technique did not improve the accuracy, but atleast it shoulder be a starting point for our stakeholder, Department of Education to proactively monitor and improve the performance of the schools. Though we could get a high accuracy on deterring the magnitude of school-performance improvement, we were able to classify the improvement.

Classification of School-performance improvement.

We used SVM, KNN, DT and logistic classifier to determine whether the schools will improve or not. We had to change our target variable to be discrete variable instead of continues as seen on the original school performance data.

Model Performance

In this section we detail and map the methods used the problem statement. We show results for each objective, and what technique was used to achieve them.

1. **Problem:** Identify the magnitude of school-performance improvement.

Solution: Use supervised learning linear models. Use the training data to Fit and predict the target variable (*Percent_2019_performance*). Use 2019 data to test the model.

	Linear Regression	Ridge	Random forest	XGBOOSt
R-Squared	**54.3**	**54.7**	**57**	**57**
RMSE	**12.04**	**12.04**	**12.04**	**12.04**

2. **Problem:** Identify the magnitude of school-performance improvement.

Solution: Use supervised learning classification models. Use the training data to Fit and predict the target variable (*Status*). Use 2019 data to test the model. NB: Status=0 if improvement <0 and 1 if improvement is >=0.

	KNN	Decision Tree	SVM	Logistic
F1-score	**76%**	**71%**	**75%**	**74**
Jaccard index	0.75	0.71	0.75	0.76

And the log losses for logistic classifier were 0.51.

3. **Problem:** cluster neighbourhood schools

   Solution : use K-means algorithm to create cluster on the Map. The 3 cluster were create d using the school-performance results from above.

Results and Discussion

Our analysis shows that although there is a great number of schools that have school performance greater than 50%. More than 80% of the school in the Eastern province have pass rate more than 50%, but the question is: ***does a school performance above 50% signifies a good quality education to high school learners***? NO, according to the government policy - every child has right to quality education. And a pass rate of 50% doesn't reflect quality education rather it shows that not enough is being done in schools to ascertain the future and success of student for Tertiary institutions and business opportunities. We have also noted from the EDM that performance is directly proportional to Quintile, this means that the student in poor community continue to fail while student coming from high income neighbourhood are excelling. Again, is this the fault with education system or failure of teaching and learning medium? Unfortunately, we don't have data about the teacher, their education level and their expertise.

CONCLUSION

Purpose of this project was to identify investigate and explore the relationship between the school neighbourhood and school performance. The necessary data was sourced and cleaned. Various linear models and classifiers we used to achieve the project objective. The prediction accuracy

were discussed and mapped to address the problems statement. The results showed that the classifier model had an average of 74% which meant better chance of classifying the improvement, while linear model had an average of 54% which just more than half of the schools we could determine the improvement magnitude for. This project serves as starting point for a proactive improvement of school performance in EC.

FUTURE ACTION

With the pandemics like COVID19, the use of data to determine and assess the quality of education high school student receive is important. With more that and resources, we recommend the use the above model with teacher's information included and school subject info included, the number of students to write in 2020, so that Tertiary institutions can be aware of what type of first year student they are dealing with. Furthermore, the department can align the curriculum and lost academic days due COvid19 lockdowns, using historic data to predict the 2020 results. SO that schools that experience issue can be dealt with before the final exams.