CSE 5160 Summer 2021
Group Final Project
Data Analysis on Student Performance

Group 5
Gil Alvarez
Christopher Magnuson
Takenori Tsuruga

Background:

- Student success in Junior High and High school can have a large impact on future success in college.
- Can attributes such as parent's job, income, access to tutoring services, access to internet help us predict success in classes.
- Our group consulted a data set provided by Professor Paulo Cortez from the University of Minho in Portugal that followed students at two different secondary schools.
- Students were chosen from these two schools that were enrolled in either Mathematics or Portuguese.
- Students were give a survey that asked several questions, father's job, mother's job, number of past class failures, etc.

Research Question

- Given the following dataset, what corresponding attributes given a strong correlation increases the accuracy of predicting a student's success rate?
- What training method out of the models we implemented are suitable for each datasets?
- For each model, which predictor setup performed better? all or selected.
- For the training models comes with parameter, what value was found optimal for each case?

Attribute	Description	Domain	Note
school	student's school	binary	Gabriel Pereira or Mousinho da Silveira
sex	student's sex	binary	Female or male
age	student's age	numeric	From 15 to 22
address	student's home address	binary	urban or rural
famsize	Family size	binary	≤ 3 or > 3
Pstatus	Parent's cohabitation status	binary	Living together or apart
Medu	mother's education	numeric	From 0 to 4
Fedu	father's education	numeric	from 0 to 4
Mjob	Mother's job	nominal	teacher, health care, civil services, at home or other
Fjob	Father's job	nominal	teacher, health care, civil services, at home or other
reason	Reason to choose this school	nominal	close to home, school reputation, course preference or other
guardian	Student's guardian	nominal	mother, father or other
traveltime	Home to school travel time	numeric	1: < 15 m, 2: 15 to 30 m, 3: 30 m to 1 h, 4: > 1 h
studytime	Weekly study time	numeric	1: < 2 h, 2: 2 to 5 h, 3: 5 to 10 h, 4: > 10 h
failures	Number of past class failures	numeric	n if 1 ≤ n < 3, else 4
schoolsup	Extra educational school support	binary	yes or no

Attribute	Description	Domain	Note
famsup	extra educational support	binary	yes or no
paid	Extra paid classes within course	binary	yes or no
activities	extra-curricular activities	binary	yes or no
nursery	Attended nursery school	binary	yes or no
higher	Wants to take higher education	binary	yes or no
internet	Internet access at home	binary	yes or no
romantic	with a romantic relationship	binary	yes or no
famrel	Quality of family relationships	binary	yes or no
freetime	free time after school	numeric	from 1 – very low to 5 – very high
goout	going out with friends	numeric	from 1 – very low to 5 – very high
Dalc	workday alcohol consumption	numeric	from 1 – very low to 5 – very high
Walc	weekend alcohol consumption	numeric	from 1 – very low to 5 – very high
health	current health status	numeric	from 1 – very bad to 5 – very good
absences	number of school absences	numeric	from 0 to 93
G1	first period grade	numeric	from 0 to 20
G2	second period grade	numeric	from 0 to 20
G3	final grade	numeric	from 0 to 20

Data Analysis Methods:

Approach:

2 Datasets
- Mathematics (395 samples 33 attributes)
- Portuguese (649 samples 33 attributes)

Predictor for Training
- All predictors
- Selected predictors

Sampling
- K-fold cross validation with K=10

Analysis and Training Models
- Regression
  - Multiple Linear Regression
  - K-Nearest Neighbors
  - Support Vector Machine
    - Linear Kernel
    - Radial Kernel
- Classification (Fail/Pass bi-classification)
  - Linear Discriminant Analysis
  - K-Nearest Neighbors
  - Support Vector Machine
    - Linear Kernel
    - Radial Kernel

Data Analysis Methods:

Data Preprocessing:

Data Structure Conversion
- Binary
  character -> factor
- Nominal
  character -> factor

Standardization on Numeric
- Standard deviation of 1
- Center the mean at 0

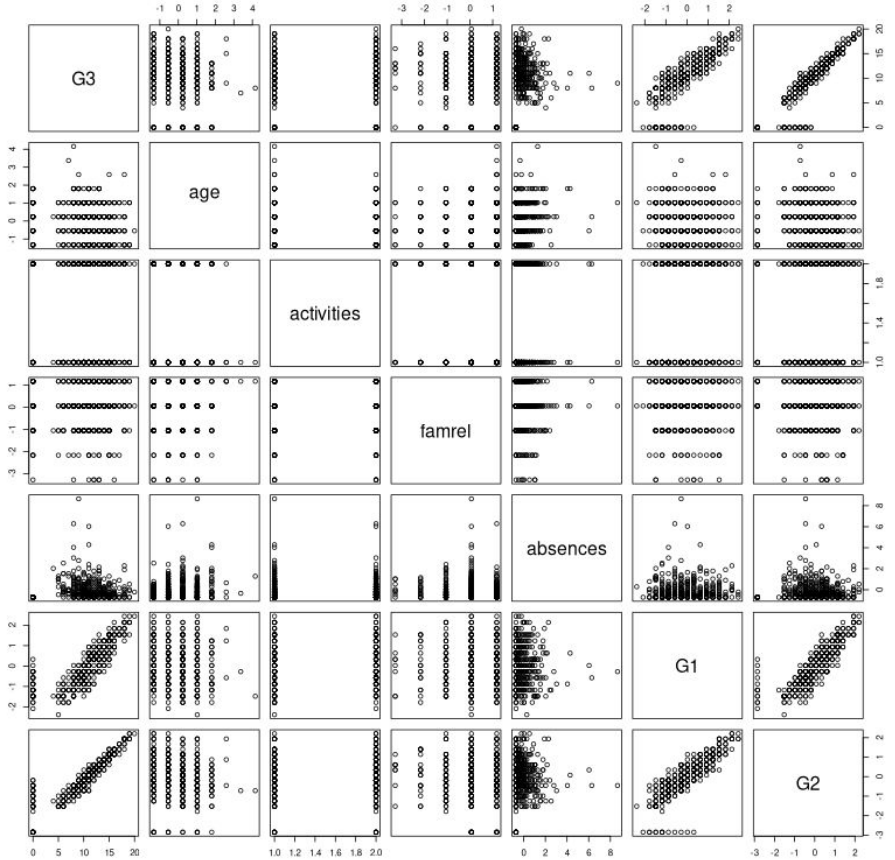
Dataset for Classification
- Duplicate each dataset
- Final Grade
  numeric -> factor
  1-20 -> Fail/Pass

Selection of Predictors:

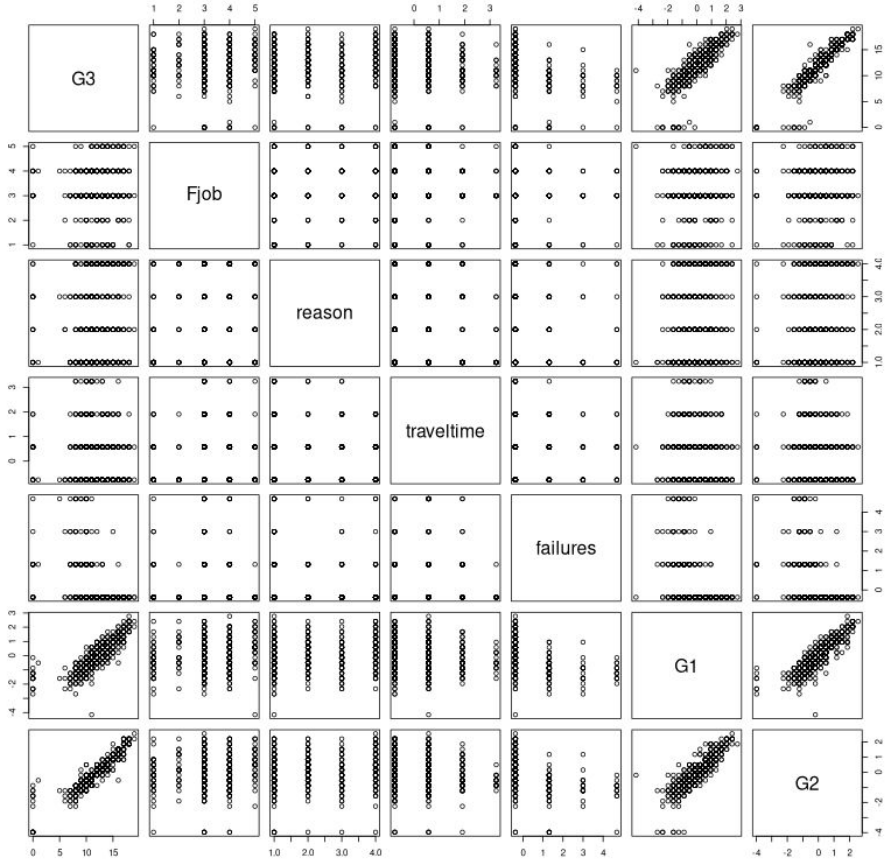
- Run lm() on each dataset
- Run summary() to get the p-value of each predictor from output
- p-values shown on right table
- Select attribute
  p-value < 0.1
- Mathematics
  age, activities, famrel, absences
  G1, G2
- Portuguese
  Fjob, reason, traveltime, failures
  G1, G2

Coefficients	Mat P-Value	Por P-Value
(Intercept)	< 2e-16	< 2e-16
SchoolMS	.190485	.121992
sexM	.455805	.298423
age	**.086380**	.553208
addressU	.699922	.351565
famsizeLE3	.872128	.892197
PstatusT	.703875	.549055
Medu	.387859	.196799
Fedu	.298974	.442773
Mjobhealth	.777796	.292379
Mjobother	.823565	.510720
Mjobservices	.898973	.324808
Mjobteacher	.956522	.348232
Fjobhealth	.619871	.208189
Fjobother	.860945	.114544
Fjobservices	.514130	**.036457**
FjobTeacher	.851907	**.085958**
reasonhome	.415123	.555479
reasonother	.419120	**.036251**
reasonreputation	.629335	.226584
guardianmother	.439046	.840252
guardianother	.988710	.383539
traveltime	.539170	**.063667**
studytime	.437667	.453569
failures	.319399	**.010254**
schoolsupyes	.154043	.287969
famsupyes	.430710	.377230
paidyes	.733211	.376663
activitiesyes	**.093774**	.908275
nurseryyes	.381518	.452553
higheryes	.651919	.256285
internetyes	.615679	.511152
romanticyes	.216572	.696483
famrel	**.001912**	.770469
freetime	.670021	.342694
goout	.909224	.708033
Dalc	.227741	.469977
Walc	.124966	.760521
health	.400259	.129064
absences	**.000698**	.247198
G1	**.002645**	**.000626**
G2	**< 2e-16**	**< 2e-16**

Mathematics Selected Predictor



Portugese Selected Predictor



Data Analysis Methods:

Regression Analysis Workflow:

Model Training
- Mathematics with all predictors
- Mathematics with selected predictors
- Portuguese with all predictors
- Portuguese with selected predictors
- Various model parameters if any

Plots
- Scatter Plots of each trained model

Table
- Trained model details
  - Model Parameter(s)
  - RMSE, etc

Classification Analysis Workflow:

Model Training
- Same as Regression

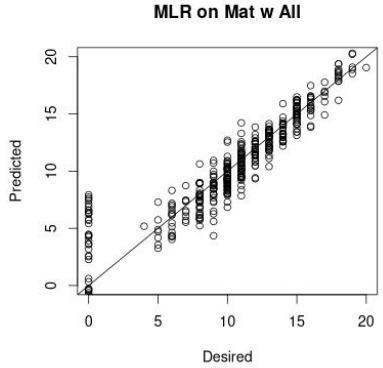
Table
- Trained model details
  - Model Parameter(s)
  - Accuracy, ROC, etc
- Confusion matrices

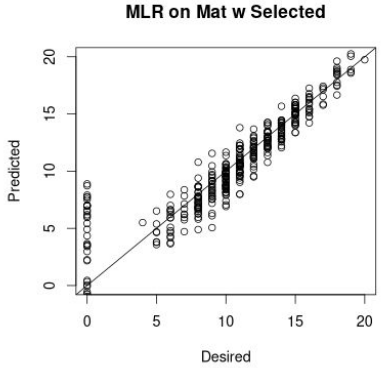
Select Direction for Next Step of Analysis
- All or Selected predictors
- Accuracy or ROC based performance

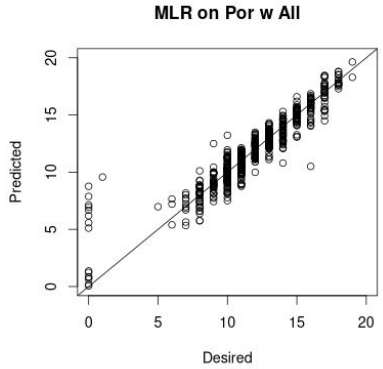
Plots
- ROC curves with optimal thresholds
- Closest point to left top corner

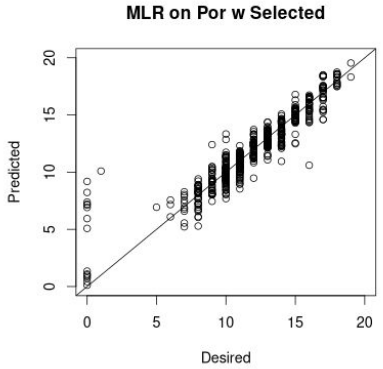
Final confusion matrices

Multiple Linear Regression: Regression









MLR: Regression

	RMSE	Rsquared	MAE
Mat w All	1.997856	.809855	1.336831
Mat w Sel	1.866286	.8310588	1.185114
Por w All	1.271458	.8473582	.8177206
Por w Sel	1.240351	.8560713	.7914463

While the scatter plots shows similar results, RMSE from the above table shows improvement of performance models, with selected predictors.

Linear Discriminant Analysis: Classification

LDA: Classification Accuracy ROC

	Accuracy	Kappa	ROC	Sens	Spec
Mat w All	.8786538	.7181369	.95778	.7846154	.9246439
Mat w Sel	.9088462	.7908273	.9779969	.8461538	.9396011
Por w All	.9028846	.5854038	.9319158	.58	.9617172
Por w Sel	.9090865	.6008262	.9496599	.57	.9708754

Both Mat and Por shows improved performance with selected predictors over all predictors in terms of both accuracy and ROC measurements.

LDA: Confusion Matrices Accuracy ROC

Mat w All

	Fail	Pass	Accuracy	.9418
Fail	117	10	Sensitivity	.9000
Pass	13	255	Specificity	.9623

Mat w Sel

	Fail	Pass	Accuracy	.919
Fail	111	13	Sensitivity	.8538
Pass	19	252	Specificity	.9509

Por w All

	Fail	Pass	Accuracy	.9307
Fail	69	14	Sensitivity	.6900
Pass	31	535	Specificity	.9745

Por w Sel

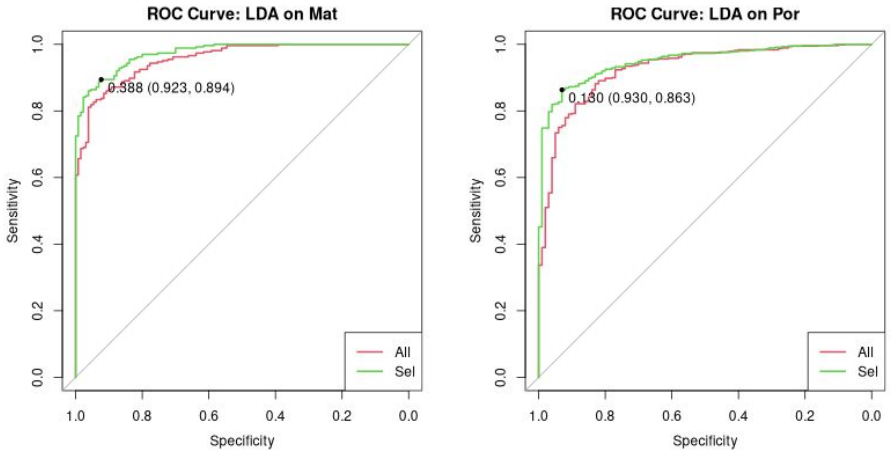
	Fail	Pass	Accuracy	.9137
Fail	59	15	Sensitivity	.59
Pass	41	534	Specificity	.97268

Accuracy rate from the confusion matrices indicates the opposite.

Our pick for next step

- Selected predictors

Linear Discriminant Analysis:



Mat w Sel

	Fail	Pass	Accuracy	.919
Fail	111	13	Sensitivity	.8538
Pass	19	252	Specificity	.9509

Por w Sel

	Fail	Pass	Accuracy	.9137
Fail	59	15	Sensitivity	.59
Pass	41	534	Specificity	.97268

LDA: Final Confusion Matrices

Mat w Sel (Thres=0.388)

	Fail	Pass	Accuracy	.9013
Fail	120	29	Sensitivity	.9231
Pass	10	236	Specificity	.8906

Por w Sel (Thres=0.13)

	Fail	Pass	Accuracy	.8798
Fail	92	70	Sensitivity	.9200
Pass	8	479	Specificity	.8725

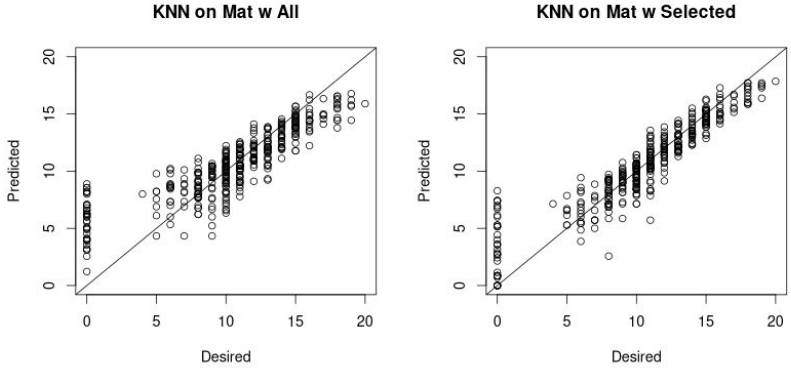
ROC curves for both dataset displays advantage of selected predictor models.

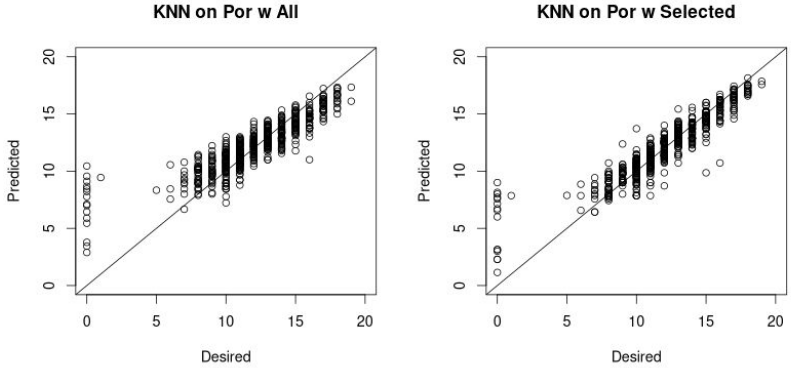
Optimal Threshold: Mat = 0.388
Por = 0.13

Application of threshold greatly improved sensitivity in both cases with trade of of loss in accuracy and specificity, as expected.

With Portuguese, sensitivity increased from 59% to 92% which is remarkable.

K-Nearest Neighbor: Regression





K-Nearest Neighbors: Regression

	k	RMSE	Rsquared	MAE
Mat w All	9	2.718487	.7056206	1.926936
Mat w Sel	7	1.966059	.8243143	1.357927
Por w All	9	1.867822	.7107004	1.289693
Por w Sel	7	1.517768	.7941841	.9862281

Improvement of performance can be observed with selected predictors on both scatter plots and RMSE.

K-Nearest Neighbor: Classification

KNN: Classification Accuracy

	k	Accuracy	Kappa
Mat w All	31	.8353846	.5803307
Mat w Sel	41	.8962179	.7500379
Por w All	1	.8766106	.4624003
Por w Sel	17	.9044231	.5547594

KNN: Classification ROC

	k	ROC	Sens	Spec
Mat w All	47	.9466469	.4384615	.9810541
Mat w Sel	47	.9700909	.7076923	.9700855
Por w All	45	.9233788	.10	.9945455
Por w Sel	47	.9592222	.36	.9890572

Performance evaluation with accuracy is selecting k=1 for Por w All, which is known to be overfit model.

KNN does not perform well with many predictors

KNN: Confusion Matrices Accuracy

Mat w All

	Fail	Pass	Accuracy	.8354
Fail	70	5	Sensitivity	.5385
Pass	60	260	Specificity	.9811

Mat w Sel

	Fail	Pass	Accuracy	.9038
Fail	100	8	Sensitivity	.7692
Pass	30	257	Specificity	.9698

Por w All

	Fail	Pass	Accuracy	1
Fail	100	0	Sensitivity	1
Pass	0	549	Specificity	1

Por w Sel

	Fail	Pass	Accuracy	.9153
Fail	57	12	Sensitivity	.57
Pass	43	537	Specificity	.97814

KNN: Confusion Matrices ROC

Mat w All

	Fail	Pass	Accuracy	.8177
Fail	63	5	Sensitivity	.4846
Pass	67	260	Specificity	.9811

Mat w Sel

	Fail	Pass	Accuracy	.8937
Fail	94	6	Sensitivity	.7231
Pass	36	259	Specificity	.9774

Por w All

	Fail	Pass	Accuracy	.8613
Fail	14	4	Sensitivity	.14000
Pass	86	545	Specificity	.99271

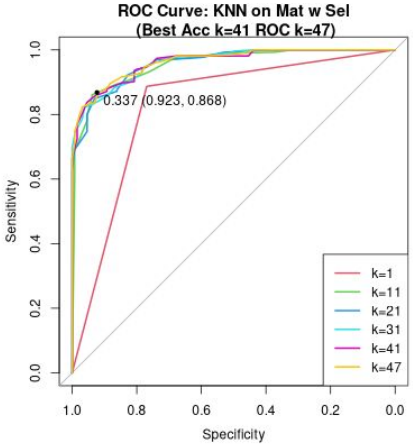
Por w Sel

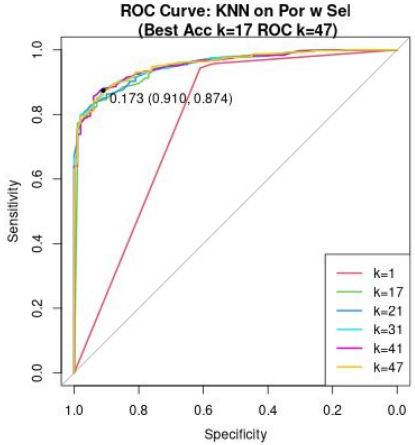
	Fail	Pass	Accuracy	.8921
Fail	37	7	Sensitivity	.37000
Pass	63	542	Specificity	.98725

Overfitness of Por w All with Accuracy is obvious from the confusion matrices

Our pick for next step -> ROC with selected predictors

K-Nearest Neighbor: Classification





ROC curses for both dataset along with the variation of the curve on different k values.

Optimal Threshold: Mat = 0.337
Por = 0.173

Mat w Sel

	Fail	Pass	Accuracy	.8937
Fail	94	6	Sensitivity	.7231
Pass	36	259	Specificity	.9774

Por w Sel

	Fail	Pass	Accuracy	.8921
Fail	37	7	Sensitivity	.37000
Pass	63	542	Specificity	.98725

KNN: Final Confusion Matrices

Mat w Sel (Thres=0.337)

	Fail	Pass	Accuracy	.881
Fail	120	37	Sensitivity	.9231
Pass	10	228	Specificity	.8604

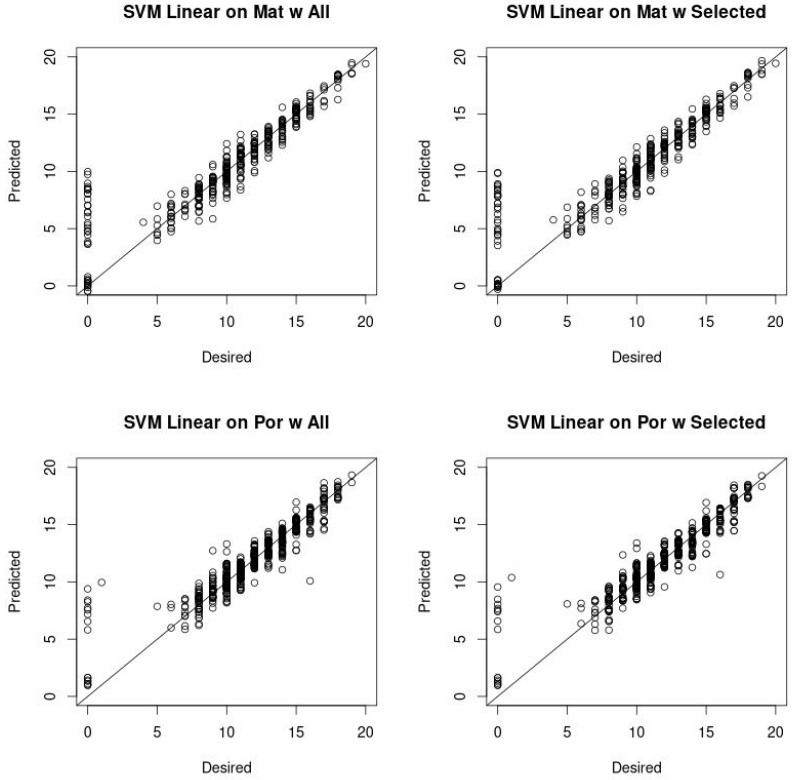
Por w Sel (Thres=0.173)

	Fail	Pass	Accuracy	.886
Fail	94	68	Sensitivity	.9400
Pass	6	481	Specificity	.8761

Trade-off loss in accuracy is managed around only 1% on both cases.

With Portuguese, sensitivity increased from 37% to 94% which is greater change than one observed with LDA.

SVM Linear kernel: Regression



SVM Linear Kernel: Regression

	C	RMSE	Rsquared	MAE	Support Vectors
Mat w All	0.1	1.982628	.8156515	1.100968	245
Mat w Sel	10	1.938234	.8257238	1.050378	227
Por w All	10	1.256297	.8528415	.7930066	505
Por w Sel	1	1.242843	.8560689	.7844719	544

There is not much difference that can be observed from the scatter plots between all predictors and selected predictors.

In terms of RMSE, there is slight difference on Mat in favor of selected predictors.

SVM Linear kernel: Classification

SVM Linear Kernel: Classification Accuracy

	C	Accuracy	Kappa	SV
Mat w All	1	.9267308	.8324107	73
Mat w Sel	1	.9189744	.8152175	85
Por w All	0.1	.9275721	.6886938	137
Por w Sel	0.1	.9291346	.6972342	137

SVM Linear Kernel: Classification ROC

	C	ROC	Sens	Spec	SV
Mat w All	1	.9788516	.8846154	.9472934	73
Mat w Sel	1	.9794762	.8615385	.9472934	85
Por w All	0.1	.9529933	.65	.9781145	137
Por w Sel	0.1	.9677778	.66	.9781481	137

Both accuracy and ROC points the exact same C value for each case.

SVM in general are known to perform better with higher predictor dimensions.

SVM Linear Kernel: Confusion Matrices Accuracy ROC

Mat w All

	Fail	Pass	Accuracy	.9722
Fail	126	7	Sensitivity	.9692
Pass	4	258	Specificity	.9736

Mat w Sel

	Fail	Pass	Accuracy	.9266
Fail	114	13	Sensitivity	.8769
Pass	16	252	Specificity	.9509

Por w All

	Fail	Pass	Accuracy	.9492
Fail	72	5	Sensitivity	.7200
Pass	28	544	Specificity	.9909

Por w Sel

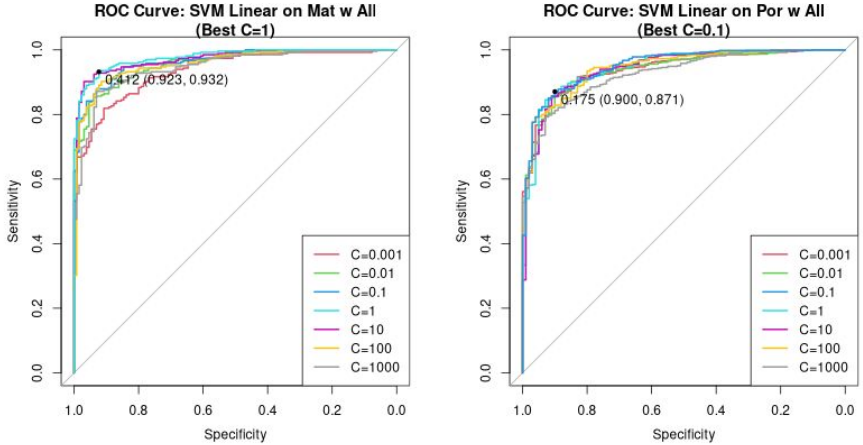
	Fail	Pass	Accuracy	.9414
Fail	76	14	Sensitivity	.7600
Pass	24	535	Specificity	.9745

In terms of accuracy, they all perform pretty well on prediction as is already before applying optimal threshold.

Trained model with all predictors perform better on the predictions above.

Our pick for next step -> all predictors

SVM Linear kernel: Classification



ROC curses for both dataset along with the variation of the curve on different C values.

Optimal Threshold: Mat = 0.412
Por = 0.175

Mat w All

	Fail	Pass	Accuracy	.9722
Fail	126	7	Sensitivity	.9692
Pass	4	258	Specificity	.9736

Por w All

	Fail	Pass	Accuracy	.9492
Fail	72	5	Sensitivity	.7200
Pass	28	544	Specificity	.9909

SVM Linear: Final Confusion Matrices

Mat w All (Thres=0.412)

	Fail	Pass	Accuracy	.9722
Fail	127	8	Sensitivity	.9769
Pass	3	257	Specificity	.9698

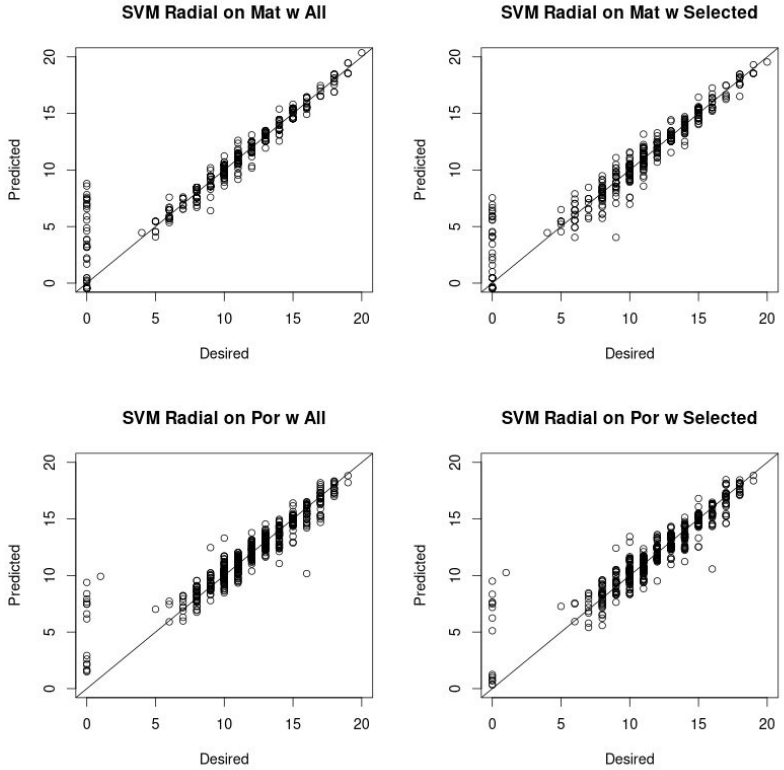
Por w All (Thres=0.175)

	Fail	Pass	Accuracy	.8752
Fail	94	75	Sensitivity	.9400
Pass	6	474	Specificity	.8634

While there is almost no change with Mat, trade off cost in accuracy on Por is noticeably big for the amount of sensitivity increase compared to previous models, LDA and KNN.

With Portuguese, sensitivity increased from 72% to 94% while losing 8% in accuracy.

SVM Radial kernel: Regression



SVM Radial Kernel: Regression

	Sigma	C	RMSE	Rsquared	MAE	Support Vectors
Mat w All	0.001	100	1.991423	.81441401	1.189906	263
Mat w Sel	0.1	10	1.700019	.86525759	1.090808	258
Por w All	0.001	10	1.272593	.85290112	.7821688	451
Por w Sel	0.001	100	1.239051	.85605230	.7806237	528

All predictors models are performing better on each datasets from the scatter plot, while RMSE suggests the opposite.

Fact above is possibly a good indication that RMSE is not really comparable between models utilizing different sets of predictors or dataset.

SVM Radial kernel: Classification

SVM Radial Kernel: Classification Accuracy

	Sigma	C	Accuracy	Kappa	Support Vectors
Mat w All	0.001	1000	.9191667	.8154401	91
Mat w Sel	0.001	1000	.9190385	.813063836	81
Por w All	0.001	100	.9275481	.69356758	133
Por w Sel	0.01	10	.9337500	.726084759	127

SVM Radial Kernel: Classification ROC

	Sigma	C	ROC	Sens	Spec	Support Vectors
Mat w All	0.001	1000	.9742604	.8538462	.9511396	91
Mat w Sel	0.001	1000	.9803528	.8461538	.9548433	81
Por w All	0.01	10	.9553333	.64	.9690236	202
Por w Sel	0.001	100	.9682357	.67	.9745118	125

Both accuracy and ROC points the exact same C and sigma value for Mat case, but different combinations for the Por.

SVM in general are known to perform better with higher predictor dimensions.

SVM Radial Kernel: Confusion Matrices Accuracy

Mat w All

	Fail	Pass	Accuracy	.9949
Fail	129	1	Sensitivity	.9923
Pass	1	264	Specificity	.9962

Mat w Sel

	Fail	Pass	Accuracy	.9241
Fail	110	10	Sensitivity	.8462
Pass	20	255	Specificity	.9623

Por w All

	Fail	Pass	Accuracy	.9676
Fail	81	2	Sensitivity	.8100
Pass	19	547	Specificity	.9964

Por w Sel

	Fail	Pass	Accuracy	.9538
Fail	79	9	Sensitivity	.7900
Pass	21	540	Specificity	.9836

SVM Radial Kernel: Confusion Matrices ROC

Por w All

	Fail	Pass	Accuracy	.9938
Fail	96	0	Sensitivity	.9600
Pass	4	549	Specificity	1.0

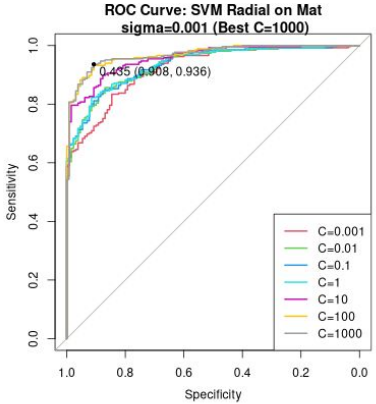
Por w Sel

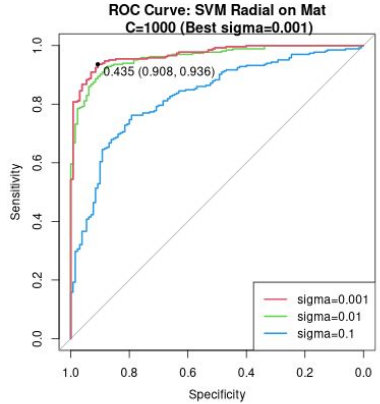
	Fail	Pass	Accuracy	.9461
Fail	75	10	Sensitivity	.7500
Pass	25	539	Specificity	.9818

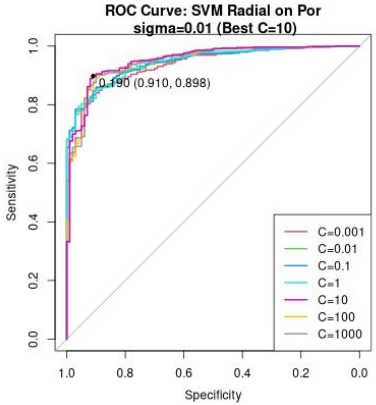
Looking through the accuracy of above predictions, we can't deny the possibility of overfit model on ones above 99%.

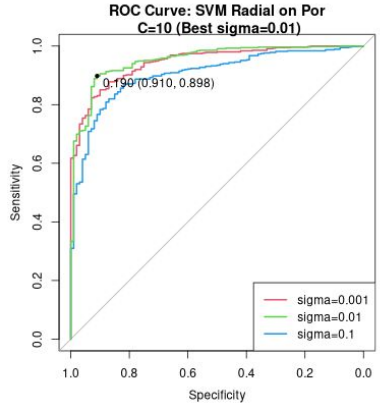
Our pick for next step -> ROC, all predictors

SVM Radial kernel: Classification









Mat w All

	Fail	Pass	Accuracy	.9949
Fail	129	1	Sensitivity	.9923
Pass	1	264	Specificity	.9962

Por w All

	Fail	Pass	Accuracy	.9938
Fail	96	0	Sensitivity	.9600
Pass	4	549	Specificity	1.0

SVM Radial: Final Confusion Matrices

Mat w All (Thres=0.435)

	Fail	Pass	Accuracy	.9949
Fail	129	1	Sensitivity	.9923
Pass	1	264	Specificity	.9962

Por w All (Thres=0.190)

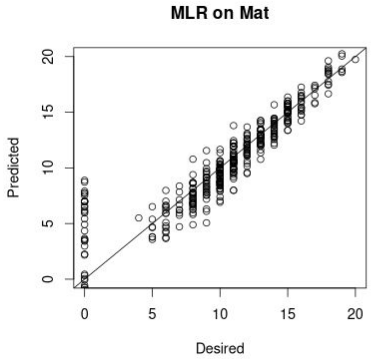
	Fail	Pass	Accuracy	.9969
Fail	100	2	Sensitivity	1.0000
Pass	0	547	Specificity	.9964

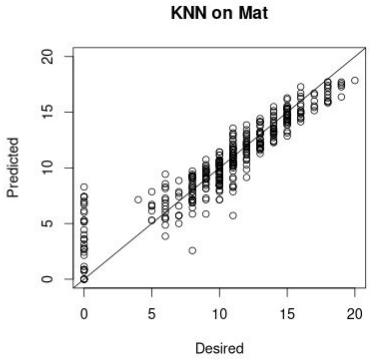
ROC curses for both dataset along with the variation of the curve on different C, sigma values.

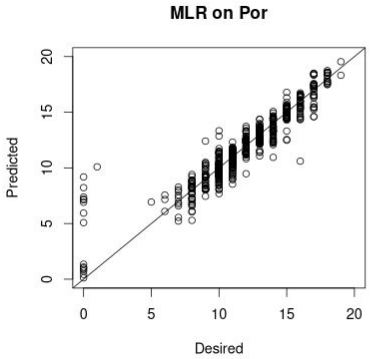
Optimal Threshold: Mat = 0.435
Por = 0.190

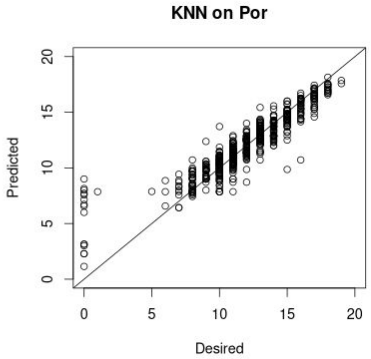
From the confusion matrices, they indicate these models are overfitted. On the other hand, from the ROC curve distance from the left top corner, there still some reasonable room remaining. This suggests the opposite possibility.

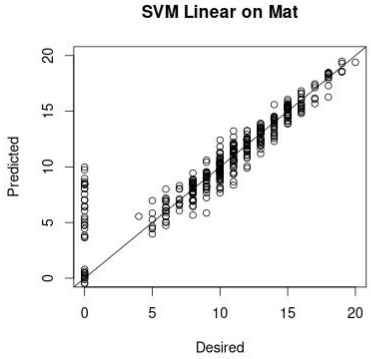
Summary of Results: Regression

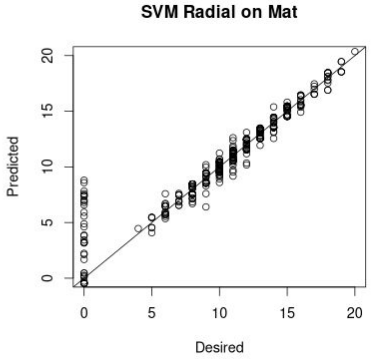


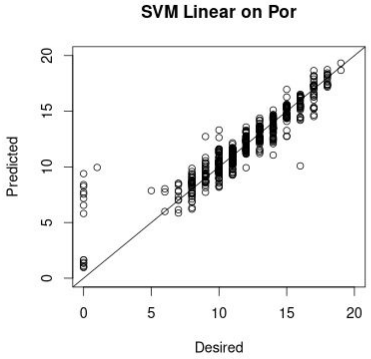


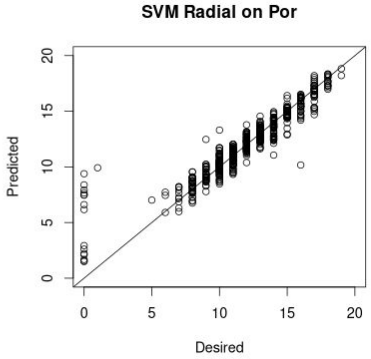




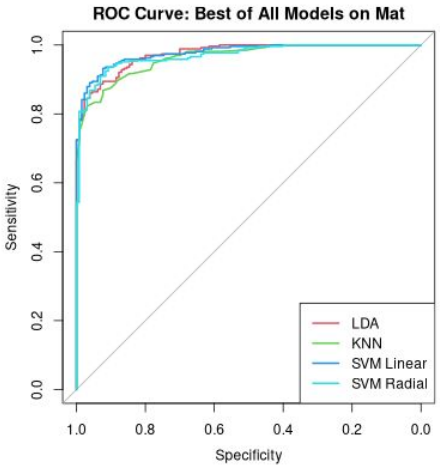


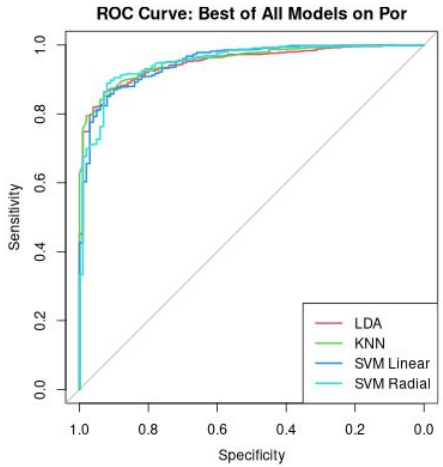






Summary of Results: Classification





Summary: Final Confusion Matrices

LDA Mat w Sel (Thres=0.388)

	Fail	Pass	Accuracy	.9013
Fail	120	29	Sensitivity	.9231
Pass	10	236	Specificity	.8906

LDA Por w Sel (Thres=0.13)

	Fail	Pass	Accuracy	.8798
Fail	92	70	Sensitivity	.9200
Pass	8	479	Specificity	.8725

KNN Mat w Sel (Thres=0.337)

	Fail	Pass	Accuracy	.881
Fail	120	37	Sensitivity	.9231
Pass	10	228	Specificity	.8604

KNN Por w Sel (Thres=0.173)

	Fail	Pass	Accuracy	.886
Fail	94	68	Sensitivity	.9400
Pass	6	481	Specificity	.8761

SVM Linear Mat w All (Thres=0.412)

	Fail	Pass	Accuracy	.9722
Fail	127	8	Sensitivity	.9769
Pass	3	257	Specificity	.9698

SVM Linear Por w All (Thres=0.175)

	Fail	Pass	Accuracy	.8752
Fail	94	75	Sensitivity	.9400
Pass	6	474	Specificity	.8634

SVM Radial Mat w All (Thres=0.435)

	Fail	Pass	Accuracy	.9949
Fail	129	1	Sensitivity	.9923
Pass	1	264	Specificity	.9962

SVM Radial Por w All (Thres=0.190)

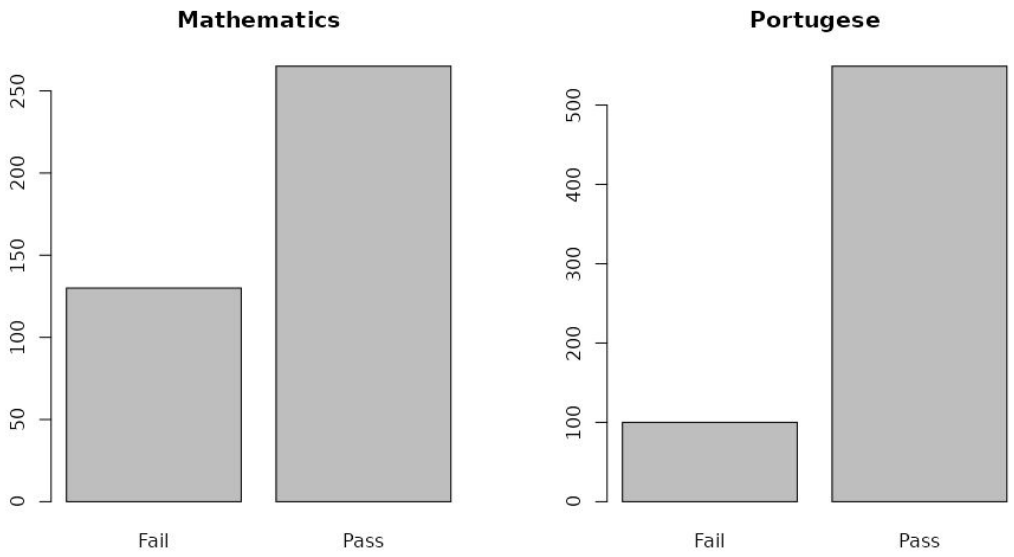
	Fail	Pass	Accuracy	.9969
Fail	100	2	Sensitivity	1.0000
Pass	0	547	Specificity	.9964

On Mat, except the KNN have a similar curve. On Por, all curves are very similar.

When curves in Mat and Por are compared, there is a trend that makes the curves on Mat closer to the left top corner.

While there are 3 models in Mat that achieves above 90% accuracy after threshold adjustment, there is only one on Por side, and has 99% accuracy. This fact indicates the model with SVM Radial on Por is overfitted.

Thoughts on Dataset Sample Distribution Ratio



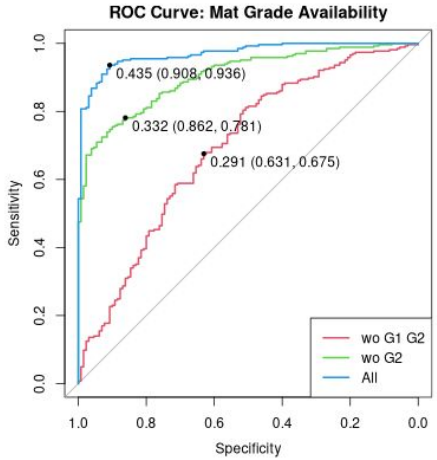
Distribution of sample per class in each dataset

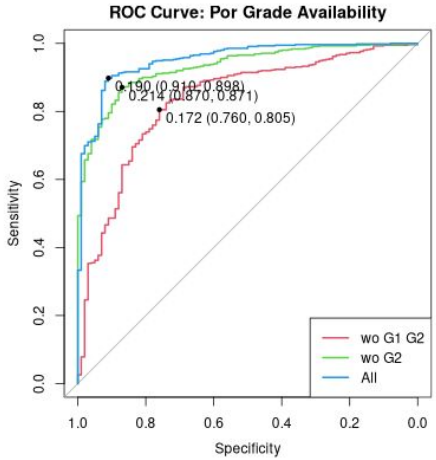
Both datasets have imbalance in distribution of each class sample.

When you consider the ratio of class distribution, while Mathematics has Fail/Pass ratio of about 1:2, Portugese has about 1:5 ratio.

The difference in ratio are significant, and it is most likely one of the major reasons why most of the trained model struggle with prediction on Portugese dataset.

Thoughts on Grade Availability





Confusion Matrices: Grade Availability

Mat wo G1 G2

	Fail	Pass	Accuracy	.7418
Fail	91	63	Sensitivity	.7000
Pass	39	202	Specificity	.7623

Mat wo G2

	Fail	Pass	Accuracy	.9848
Fail	130	6	Sensitivity	1.0000
Pass	0	259	Specificity	.9774

Mat w All

	Fail	Pass	Accuracy	.9949
Fail	129	1	Sensitivity	.9923
Pass	1	264	Specificity	.9962

Por wo G1 G2

	Fail	Pass	Accuracy	.9815
Fail	96	8	Sensitivity	9600
Pass	4	541	Specificity	.9854

Por wo G2

	Fail	Pass	Accuracy	.9337
Fail	94	37	Sensitivity	.9400
Pass	6	512	Specificity	.9326

Por w All

	Fail	Pass	Accuracy	.9969
Fail	100	2	Sensitivity	1.0000
Pass	0	547	Specificity	.9964

Above are ROC curve plots of SVM Radial kernel trained model, with variation in grade availability on both datasets.

Mat without any grade availability leads to very poor curve, while Por manage to get into somewhat acceptable shape without any grade predictor.

Once G1 become available, Mat gets into very good shape for performing good prediction accuracy.

Portuguese, on the other hand, drops accuracy of prediction from 98% to 93% with G1 grade predictor, which is an interesting effect. Still in good shape btw.

Conclusion

- The SVM Radial Kernel Model proved to be the most effective.

- Imbalance of class distribution ratio, sample size between datasets are possible causes that may account for difference of general prediction accuracy rate between two datasets.

- Grade availability affects the accuracy of prediction heavily, yet prediction model of portuguese dataset was found still feasible without any grade availability.

- In retrospective, RMSE measurement was found incomparable when the predictor number or dataset used for model trains are different sets.

Conclusion cont.

- For the model which favors small number of predictors, selected predictors case was found effective over all predictors case.

- Optimization of the classification model with ROC optimal threshold method was found to be very effective, and became a mandatory process for shaping the model into an enhanced state for real world use cases.

- Overfitting of the trained model was the raised concern throughout the project, and the safest measurement for its prevention is the use of cross validation. The method has been utilized as K-fold cross validation in this project, so we did practice the most favourable maneuver for it.

Reference

- Student Performance Data Set
  https://archive.ics.uci.edu/ml/datasets/Student+Performance

- P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
  http://www3.dsi.uminho.pt/pcortez/student.pdf

Thank you