CSE 5160

Summer 2021

Group Final Project Report

Group 5

Gil Alvarez

Christopher Magnuson

Takenori Tsuruga

Data Analysis on Student Performance Data Set

Background:

Student success in high school and junior high is an important topic that affects many countries throughout the world. Students that are not proficient in core subjects by the time they graduate can have negative impacts such as the need for remedial classes at community colleges, GED programs, and the financial burden that most students would incur to be prepared for the university level. Our group is interested to see if there are any correlations between student success and attributing factors, such as the education level of their parents, access to the internet, and family support, etc.

Background of Original Study:

In order to answer these questions we are exploring a dataset provided by Professor Paulo Cortez from the University of Minho in Portugal. This dataset followed a group of secondary school students throughout the course of an academic year. Their performance in two core subjects Mathematics and Portuguese were evaluated and several attributes such as father's job, mother's job, access to internet, etc were collected. Students were also evaluated per quarter (Q1, Q2) and their final grades for the academic year were also evaluated (Q3). Below is a list of attributes that were used in the original data set.

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 that come with parameters, what value was found optimal for each case?

Solution:

Given the fact that we have discovered attributes that have a negative value to a student's success rate we can suggest limiting negative attributes. The importance here is to use machine learning techniques to find attributes that can help increase a student's chances of success in either portuguese or mathematics. Due to the variation in datasets sizes between Mathematics and Portuguese and the difference in difficulty level from one course to the other it is possible that the attributes that give us a high prediction rate may change between the two data sets.

Dataset Attributes Table

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	Week 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 support	binary	yes or no
famsup	Family 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	numeric	from 1 – very bad to 5 – excellent
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:

Since the datasets are suitable for both regression analysis and classification, we have decided to use both approaches to analyze the data. For classification, we pick the path of bi-classification analysis with Fail or Pass, focusing on the Fail detection as the aim of the optimization. Models are trained for each dataset, mathematics and portuguese, and predictions were made for each case. While the datasets provide thirty-two predictors for training models, we made analysis with two approaches in parallel; one with all predictors, the other with few selected predictors based on the p-values. The selection of predictors were done prior to model training as part of the preparation. Details of the selection are described more in the Predictor Selection section below. Furthermore, for each regression analysis and classification, we utilized each analysis workflow to standardise the process, which will be described in the Regression Workflow and the Classification Workflow section below.

Sampling:

Following the command real data analysis case, we utilized a cross validation sampling approach with K fold configuration with K=10. This sampling methodology was applied to all the model training that took place during this project through all the training models.

Data Preprocessing:

While the datasets are provided as csv files, raw imported data into R was constructed with character and integers. In order to make use of them for processing with R, first we needed to apply data preprocessing to convert the dataset into an appropriate data structure. As they can be observed from the “Dataset Attributes Table” provided in the previous section of this report from its domain column, each attribute has one of the domains, numeric, binary, or nominal. For numeric attributes, we applied the standardization with standard deviation of 1, and centered mean to 0. For both binary and nominal domained attributes were converted into factor structure to process with R. Furthermore, we replicated the preprocessed datasets and prepared the duplicates to have bi-classified factor of “Fail” or “Pass” instead of number, which is the prediction target attribute of this project’s classification part.

Training Models:

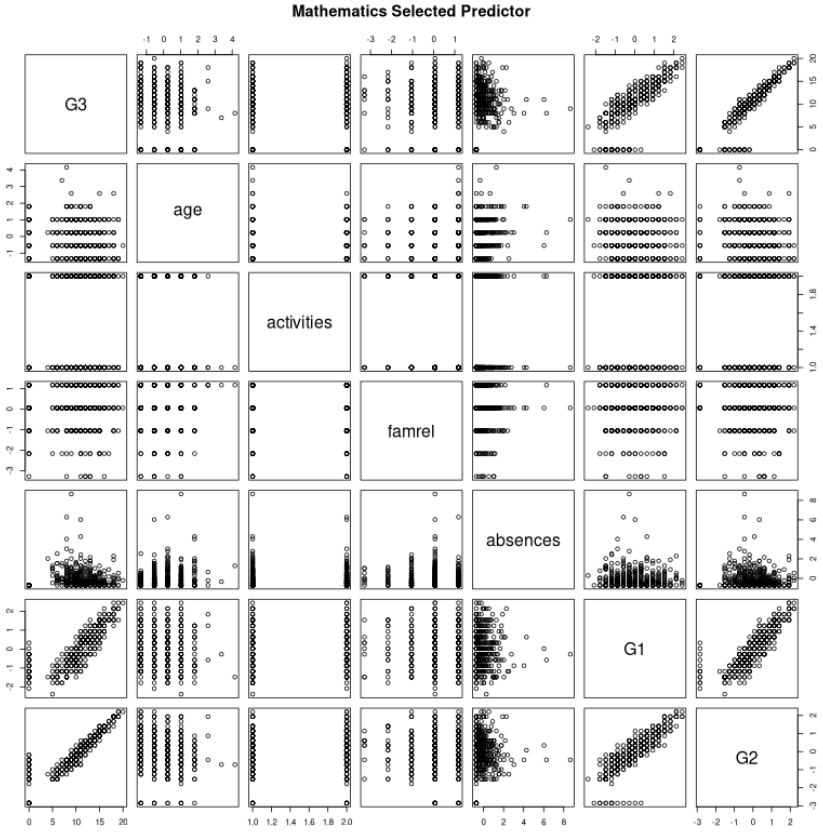
For regression analysis, we chose Multiple Linear Regression model, K-Nearest Neighbor model, Support Vector Machine with both Linear and Radial kernel. For classification, we chose Linear Discriminant Analysis model, K-Nearest Neighbor model, Support Vector Machine with both Linear and Radial Kernel.

Predictor Selection:

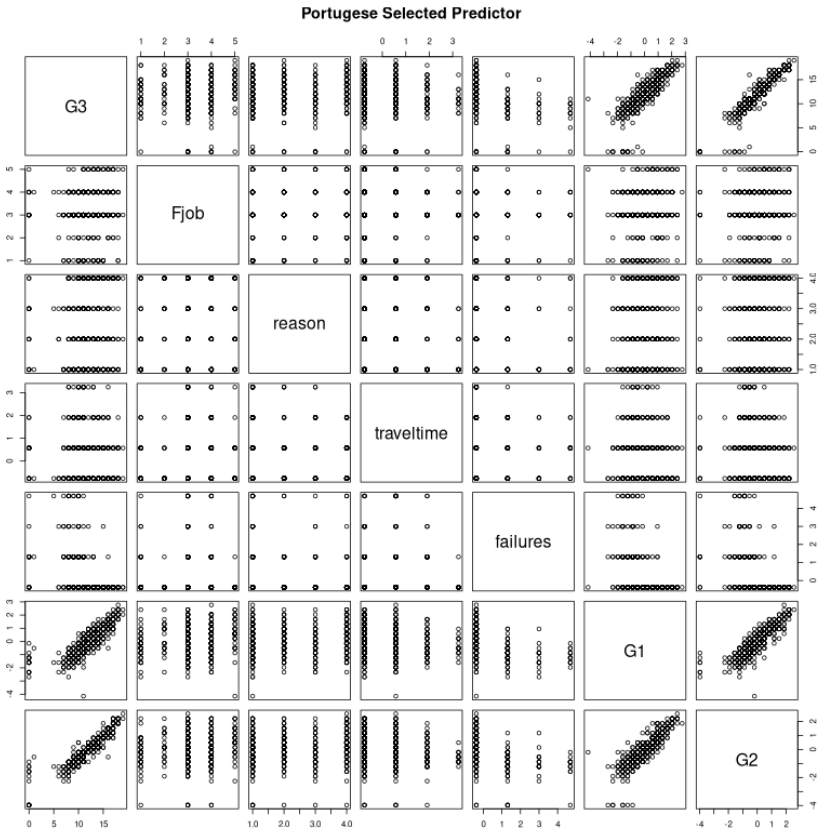
In order to determine the most influential predictors for each dataset, we utilized the lm() function with all predictors to output the p-value of each datasets which are shown in the table below.

Coefficients	Mathematics P-Value	Portuguese 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**

From the above result output from summary() function of the lm model, we decided to pick the predictors which are less than 0.1 in p-value for the selected predictor analysis. Specifically, age, activities, famrel, absences, G1, and G2 for the mathematics dataset, Fjob, reason, traveltime, failures, G1, and G2 for the portuguese dataset..



We can observe strong linear correlation of G1 and G2 to G3, moderate correlation of “famrel” and absences to G3.



Again we can observe a strong linear correlation of G1 and G2 to G3, and a moderate correlation of “failures” and “travel time” to G3

Regression Analysis Workflow:

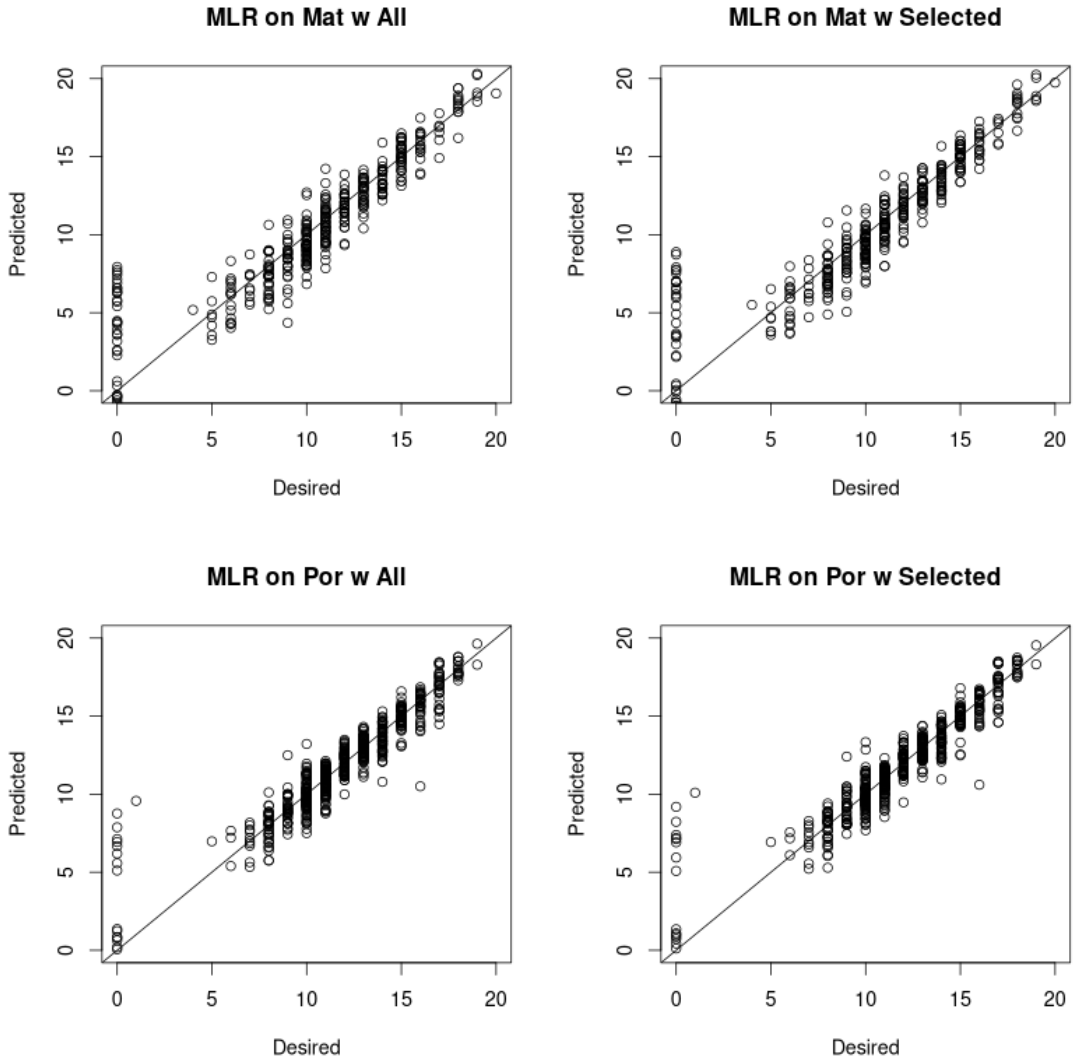
First we train the model against each dataset: mathematics and portuguese, with all or selected predictors, and display the prediction results in both scatter plots and RMSE, then make analysis on the outcome. Models which have parameter(s) to specify will explore a range of values to determine the best value for the case.

Classification Analysis Workflow:

Same as the regression case, we first train the model against each datasets with each all or selected predictors, display the prediction results in both confusion matrices, accuracy, and ROC, then make analysis on the outcome. Models which have parameter(s) to specify will explore a range of values to determine the best value for the case. For determination of the best parameter in classification, there are many methods to employ, but ROC is the commonly used metric for bi-classification. We will prioritize the decision based on it over accuracy. In addition, we will make an analysis on predictor set preference, all or selected, for each particular model based on its characteristics, and proceed to the next stage of analysis. In the second stage, we first plot the ROC curve with preferred predictors, all or selected, and the best model parameter(s). For comparison purposes, other ROC curves with other model parameters are drawn on the side. The next step is to determine the optimal threshold for the optimization of the model from the ROC curve plot. There are many methods to determine the optimal threshold, but we will utilize a method to look for the closest point to the left top of the graph as we learned from the course lecture. After finding the threshold, we apply the threshold for new prediction of the model and display the confusion matrix outcome for both datasets.

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

From the scatter plots, math tends to have slightly more scattering results with all predictors compared to selected predictors. The improved result is shown in RMSE between them. For portuguese, there is no significant difference observed between different sets. The difference in RMSE is there, but miniscule.

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

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

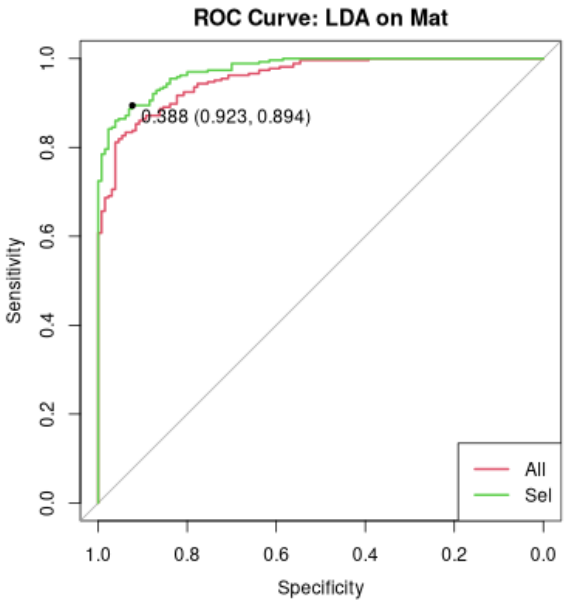
Por w All

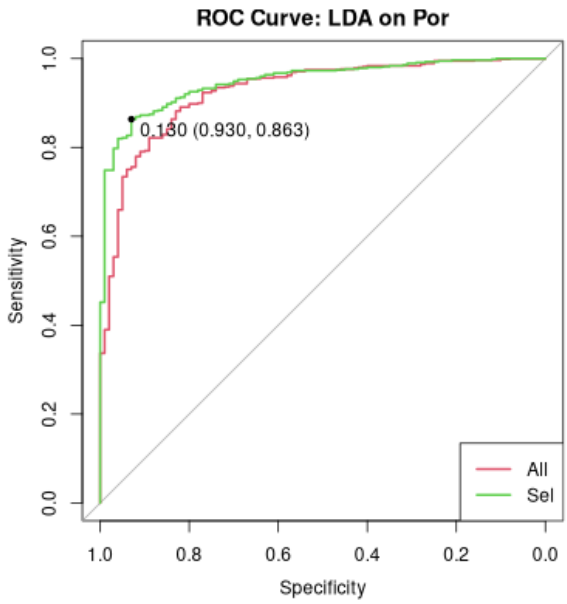
	Fail	Pass	Accuracy	.9307
Fail	69	14	Sensitivity	.6900
Pass	31	535	Specificity	.9745

Por w Sel

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

Mathematics substantially showed a better performance with the selected predictors over the rest. This can be observed by both accuracy and ROC. Ultimately, the prediction accuracy from confusion matrices shows the opposite. In the portuguese case, there is slight improvement in ROC for selected predictors, but the accuracy from the confusion matrix shows the opposite as well. However, the ROC curve from the beginning of the next step shows clear advantage of selected predictors over all predictors.





Following confusion matrices are obtained after applying the optimal threshold found from the ROC curve drawn above.

LDA: Confusion Matrices Before Threshold

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

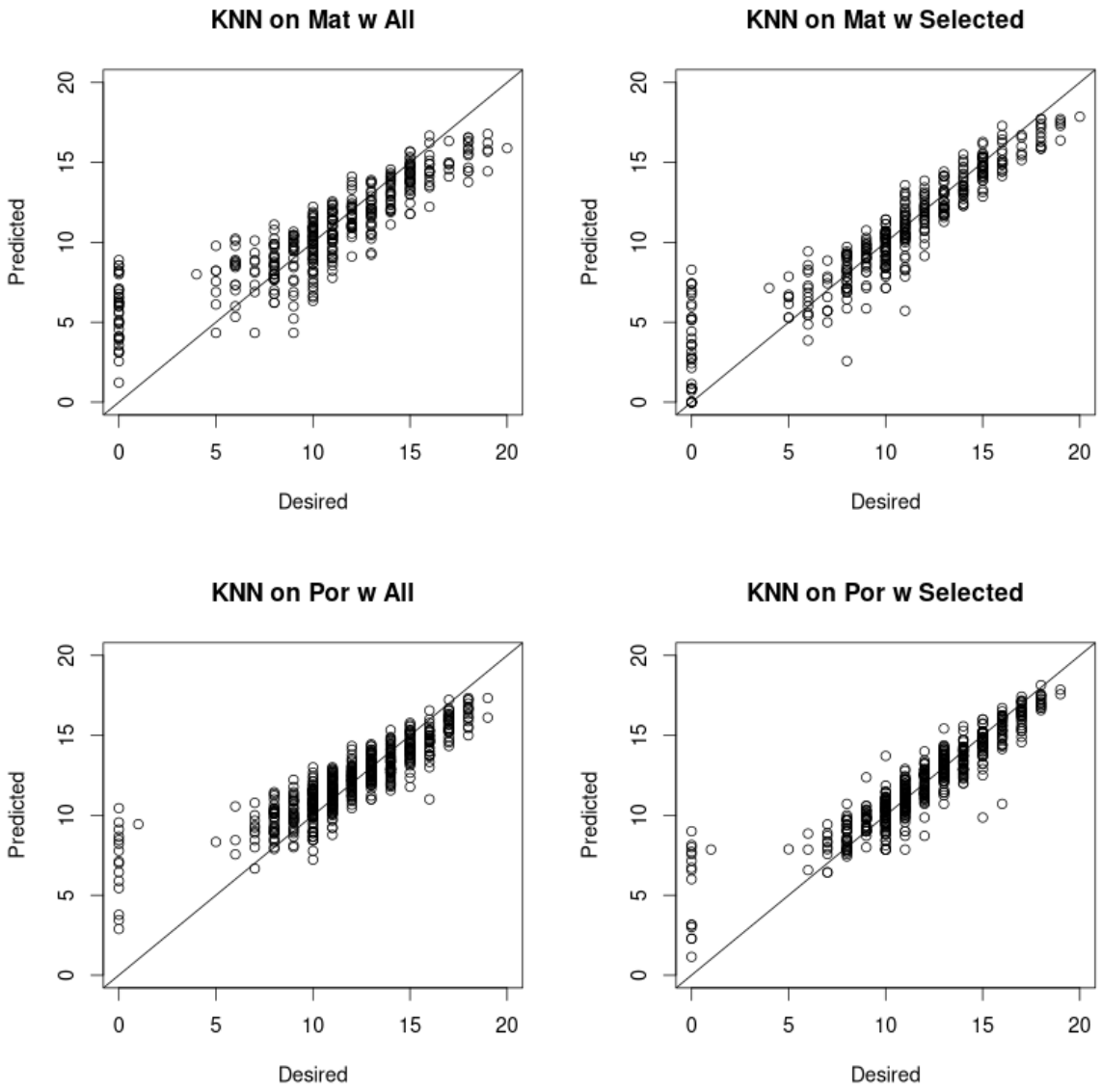
Por w Sel (Thres=0.13)

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

Applying the threshold greatly improved sensitivity (true positive rate/ Fail detection rate) while specificity and accuracy decreased as a trade off. While sensitivity was extremely low on portuguese side before applying the threshold, the outcome of applying the threshold improved vastly. It changed from unusable to a usable level.

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

The scatter plot shows that significant improvements can be observed on selected predictor variables over all predictor variables. The significance of improvement is reflected in the RMSE improvement as well. Characteristics of KNN shows that the model is not suitable for handling large numbers of predictors.

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

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

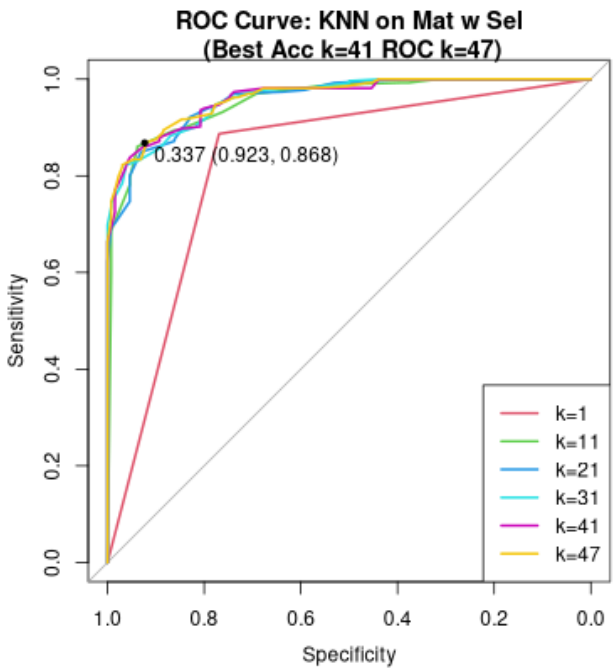
Por w All

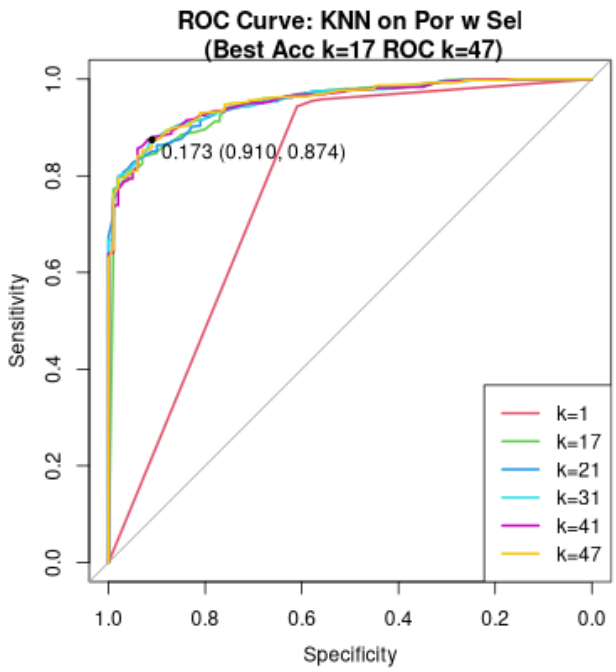
	Fail	Pass	Accuracy	.8613
Fail	14	4	Sensitivity	.14000
Pass	86	545	Specificity	.99271

Por w Sel

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

The best k value determined based on the accuracy on portuguese w all was 1, which shows over fitting results. Although it is not clear that other results from accuracy based best k value determinations are not suitable, while we are dealing with bi-classification case, we choose to respect ROC based k value determination over accuracy one. Also, as it was observed from the regression case of KNN as well, the selected predictor approach for further analysis seems to be the way to go. Favor in selected predictors over all predictors can also be observed from differences of accuracy in confusion matrices as well.





From the ROC curves plotted above, we obtained the optimal threshold for best k value curves, which is k=47 in both cases based on ROC, and the outcome of predictions with them are shown below.

KNN: Confusion Matrices Before Threshold

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

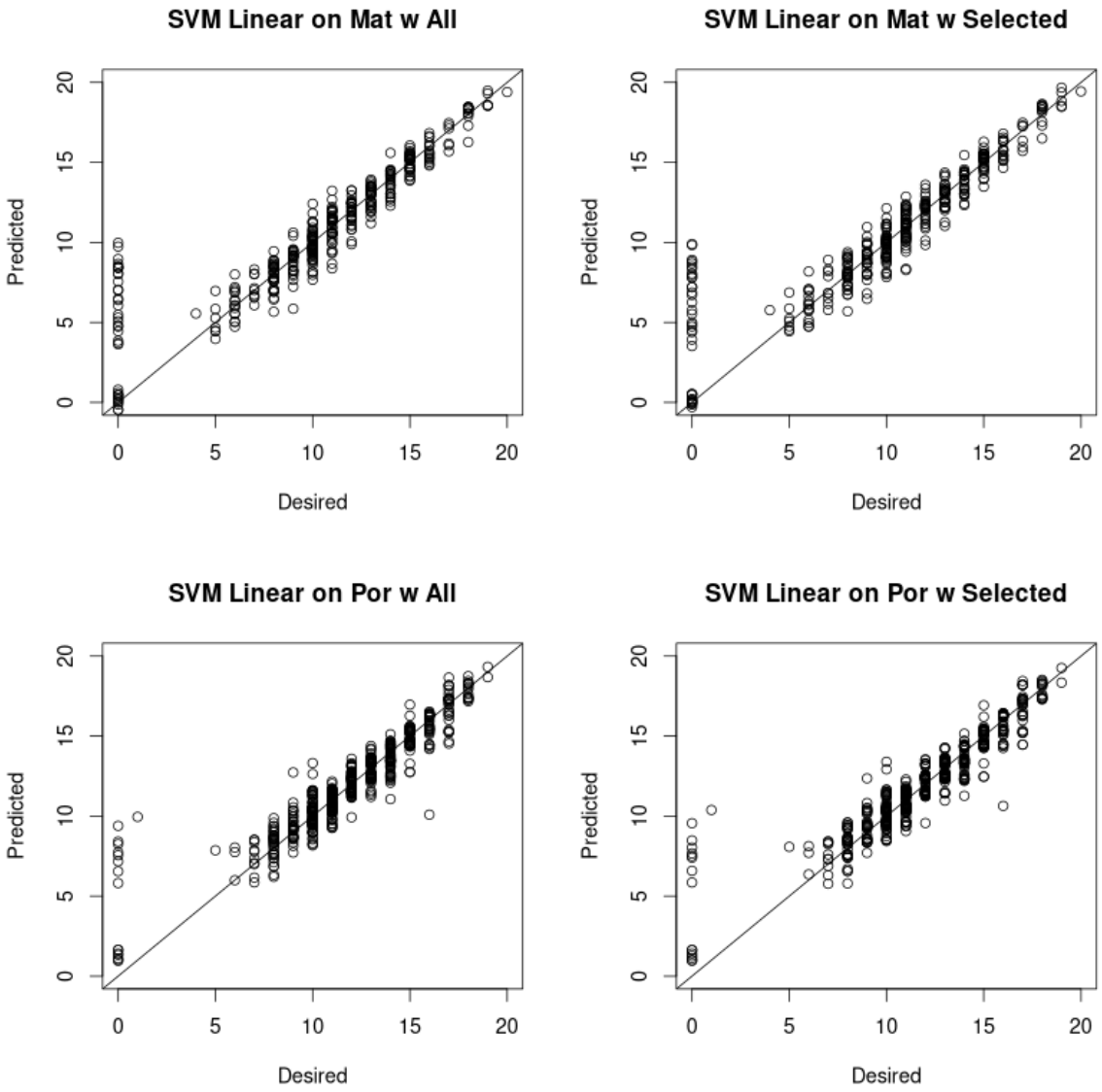
Por w Sel (Thres=0.173)

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

As expected from any threshold adjustment, we can observe the decrease in specificity and accuracy as a trade off for significant improvement on sensitivity. We can safely say they are in a somewhat usable state for the Fail detection prediction. Trade off on the accuracy is maintained around only 1% which is noteworthy.

Support Vector Machine 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

In contrast to the previous regression models, this SVM Linear Kernel model shows better performance with all predictors over selected predictors in both mathematics and portuguese cases, which is visible in both scatter plots. Considering the fact that the SVM models in general perform better in higher predictor dimensions, these results display the characteristics well.

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

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

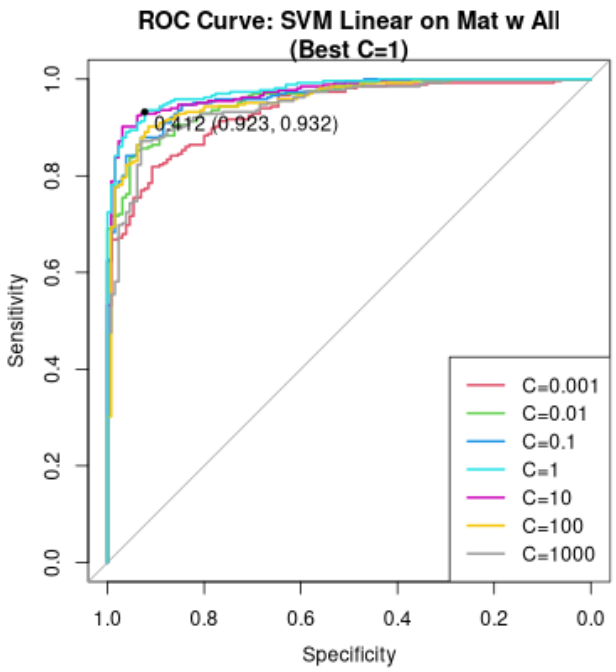
Por w All

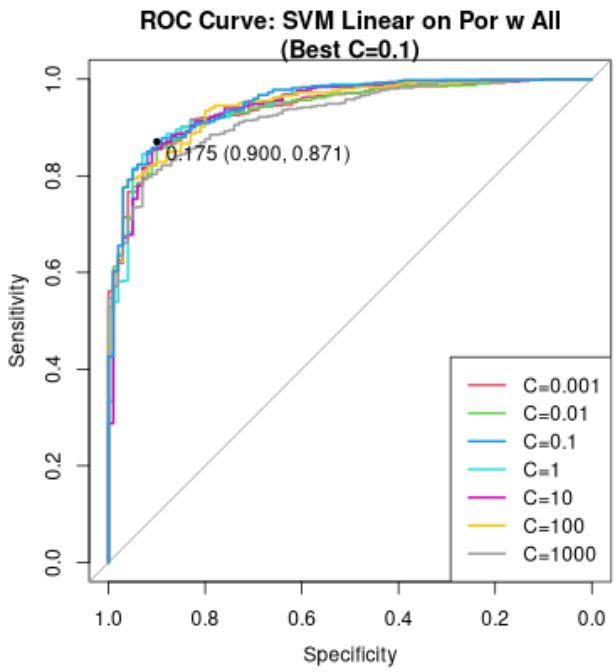
	Fail	Pass	Accuracy	.9492
Fail	72	5	Sensitivity	.7200
Pass	28	544	Specificity	.9909

Por w Sel

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

Coincidentally, accuracy and ROC both agreed on the best SVM Linear kernel parameter C in all the cases for this model. As it was mentioned in the regression analysis section of this model, it is expected to perform better on all the predictors than selected predictors, and that can be seen in accuracy from the confusion matrices. On the other hand, it is interesting that the ROC value is the only one that is in disagreement. Since we have strong reasoning for all predictors to perform better, we will proceed to the next step with the choice.





ROC curves above were plotted with variation of C value for graphical comparison purposes. The graph also displays the optimal threshold value for the best C value model for each case. Applying the acquired optimal threshold to the models, we obtained the following prediction results.

SVM Linear: Confusion Matrices Before Threshold

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

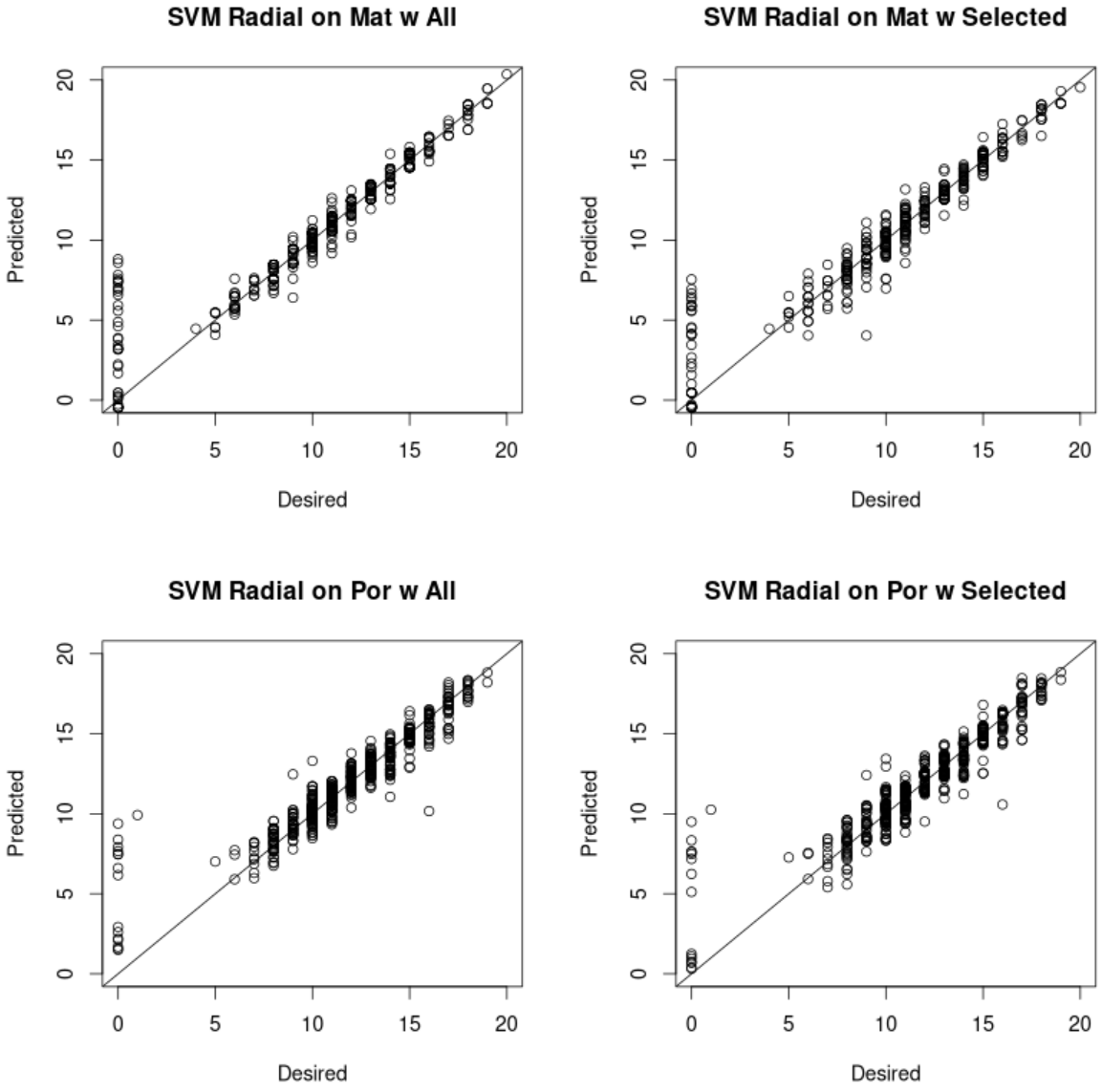
Por w All (Thres=0.175)

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

While the threshold change on mathematics was less than 0.1 and change in prediction is not that notable, improvement on portuguese model is significant in sensitivity. Of course, there are trade offs in specificity and accuracy, but as for fail detection, it can be expected to work well.

Support Vector Machine Radial Kernel:

Regression:



SVM Radial Kernel: Regression

	Sigma	C	RMSE	R Squared	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

The same trend of all predictors performing better than the selected ones from SVM Linear in general is observed here, as well from the scatter plot. Even though the performance difference shown from RMSE suggests the selected predictor is performing better, the scatter plot shows that is not the case.

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

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

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

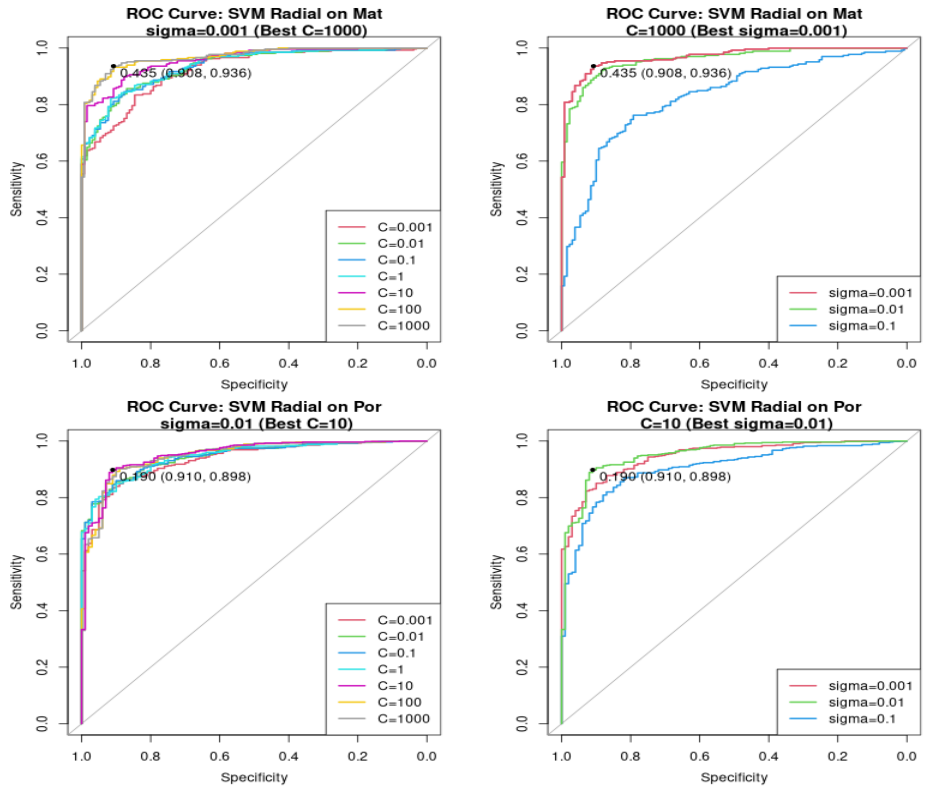
Por w All

	Fail	Pass	Accuracy	.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

Accuracy and ROC both suggested the same Radial Kernel parameter sigma and C for mathematics. They suggested different values for the portuguese, so there are variations of the predicted confusion matrices displayed above. All predictors are favored over selected predictors since it is SVM and it performs better with higher predictor dimension. This is considered the bi-classification case, in which we will pick the ROC decision over accuracy in this case.



Shown above are ROC curves displaying the variation of Radial kernel parameter sigma and C for visual comparison. Also, showing the optimal threshold value for the curve with best parameters. Obtained optimal threshold was applied to the prediction, and results are shown below in confusion matrices.

SVM Radial: Confusion Matrices Before Threshold

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

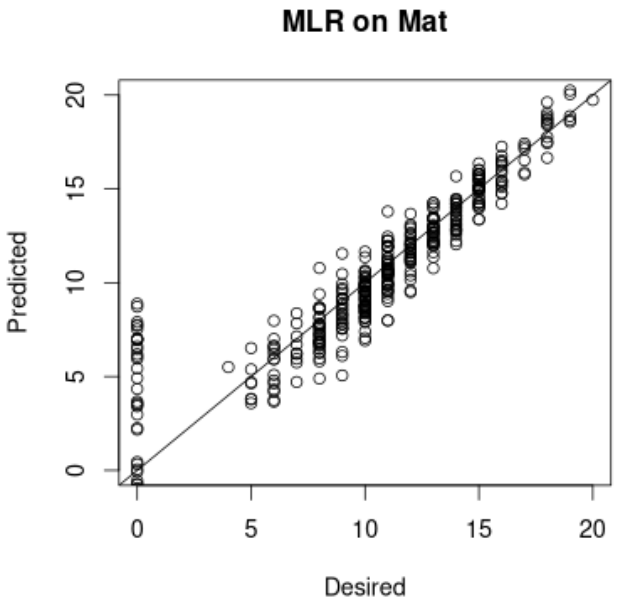
Por w All (Thres=0.190)

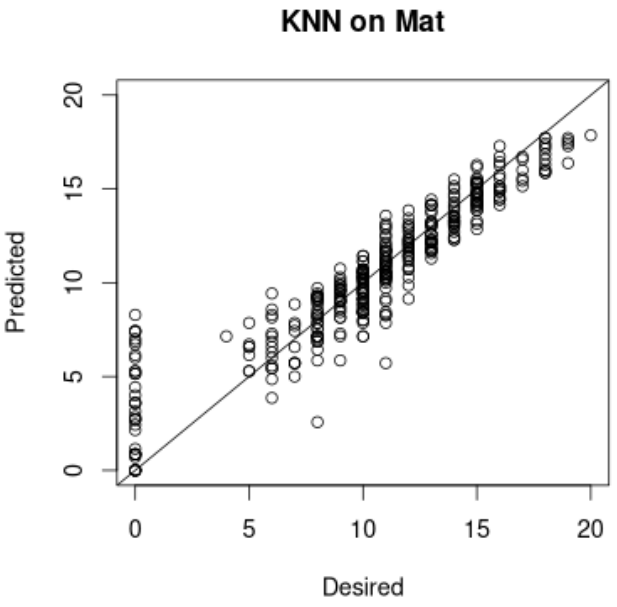
	Fail	Pass	Accuracy	.9969
Fail	100	2	Sensitivity	1.0000
Pass	0	547	Specificity	.9964

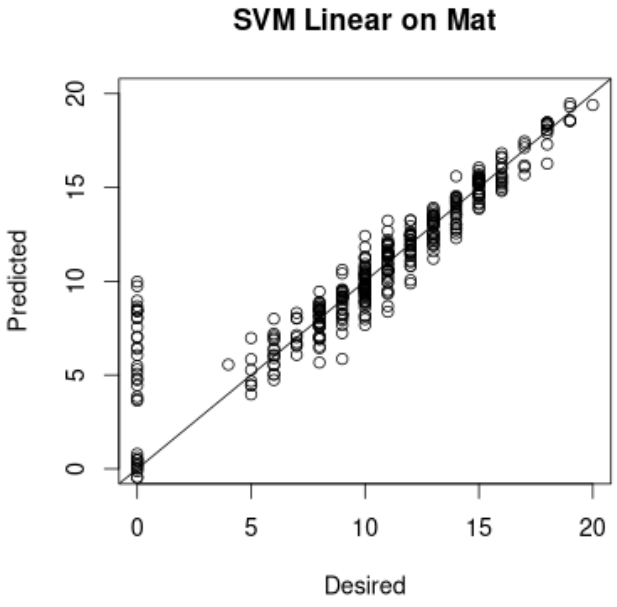
By the look of the confusion matrices above, it is possible that these are overfitted models. At the same time from the shape of the ROC curves, they are still further away from the left top corner. That suggests that these just happen to be well tuned. If these models were found to be overfitted when they are applied to the real datasets, re-tuning of C and gamma parameters will most likely help ease the tightness of fit to the training dataset.

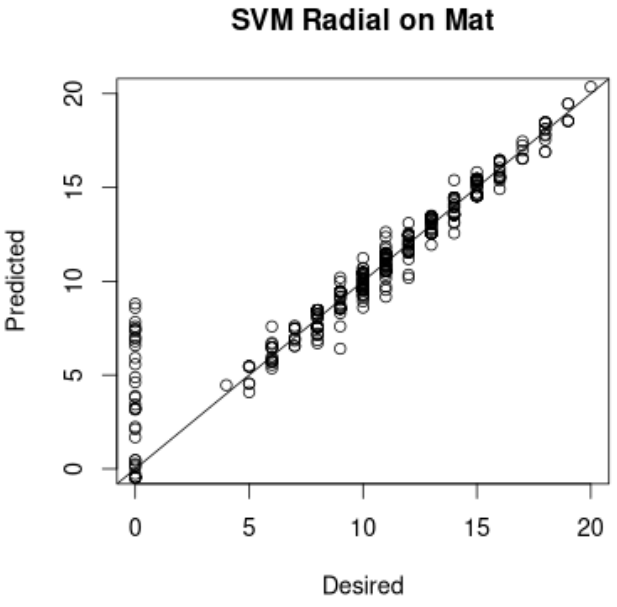
Summary of Results:

Regression:

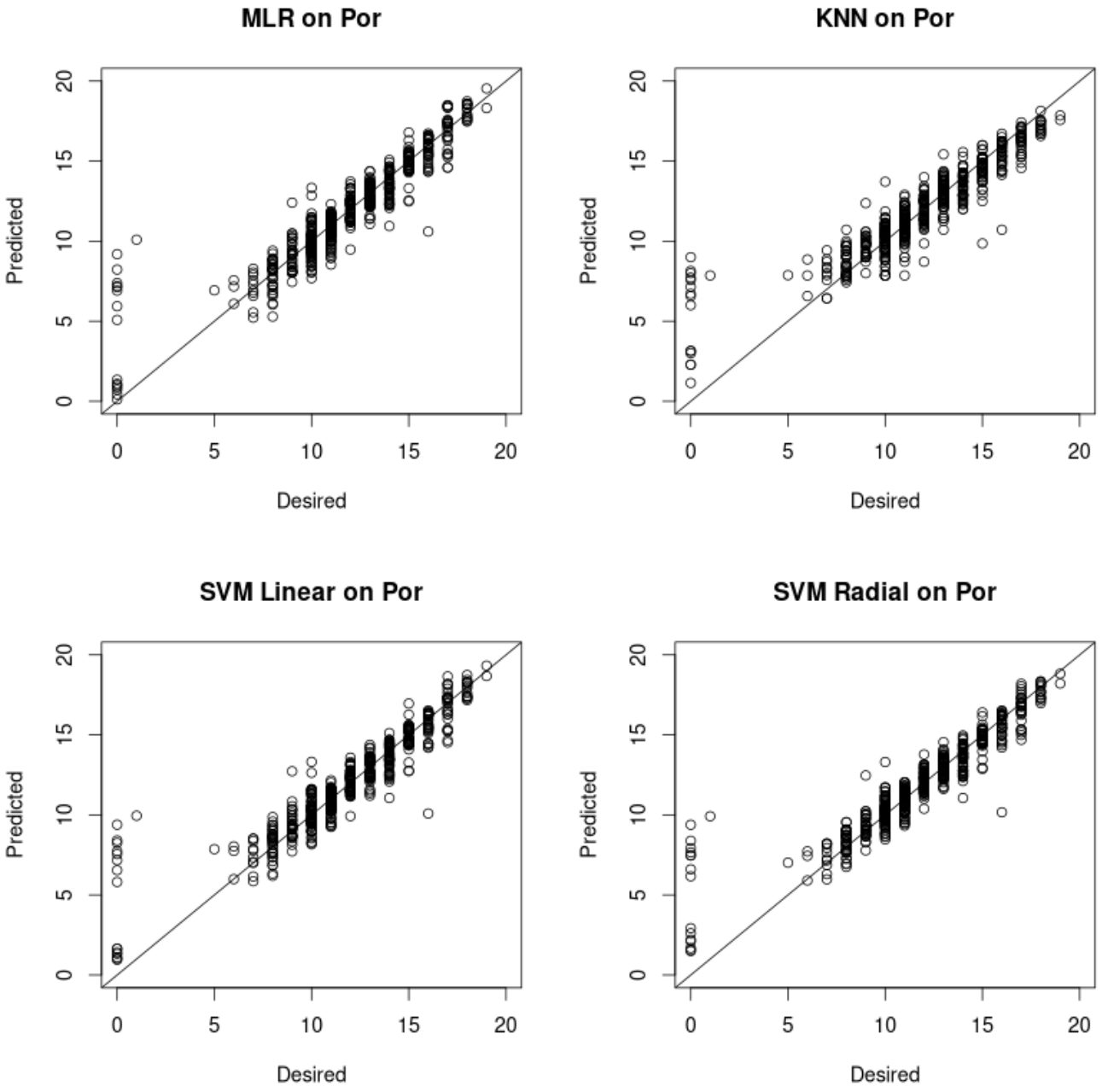






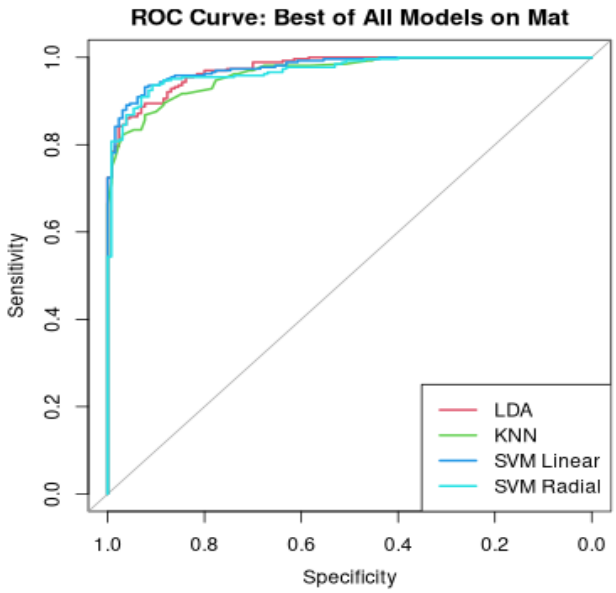


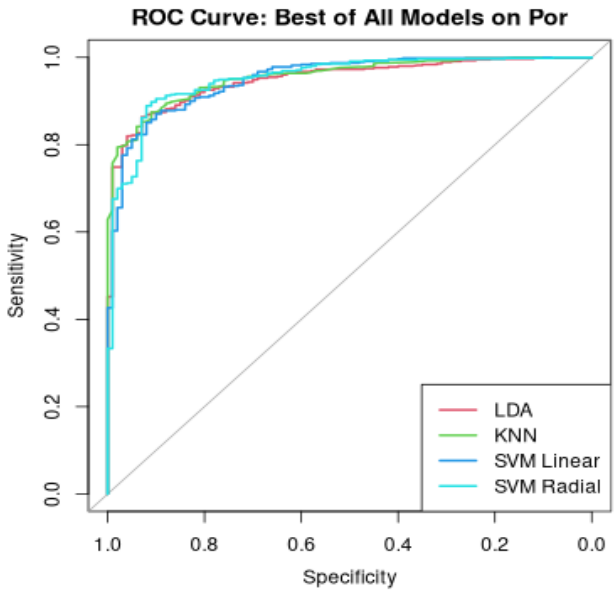
Above are the scatter plots of the best models on mathematics. In comparison of each scatter plot, good performance is shown of the SVM Radial kernel. The case states that it is significant. In contrast, the performance of KNN looks poor. To rank them, SVM Radial is the first, SVM Linear comes second, MLR third, then KNN last.



Both of the SVM plots seem to stand out in prediction performance compared to others in similar levels. We can see from the above graph that portuguese is more scattered in comparison to mathematics, which shows us that portugese is a harder case to for us to make accurate predictions on. However, the width of the scattering pattern on MLR and KNN are similar and slightly better than the mathematics case. This shows that the nature of portuguese prediction difficulty only applies to the SVM kernels. As a result, the scatter plot pattern comes very close to each other between all four of them in the portuguese case.

Classification:





The ROC curve of mathematics leads to a performance increase of both SVM models in this case. A gap of KNN performance in mathematics can be seen from the ROC curve as well. For the portuguese case, all the models are aligned in a similar manner. When you compare them to the curves from mathematics, there is a visual representation of the performance gap seen from the difference of the curves. Portuguese graphs show that they are further away from the top left corner.

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)

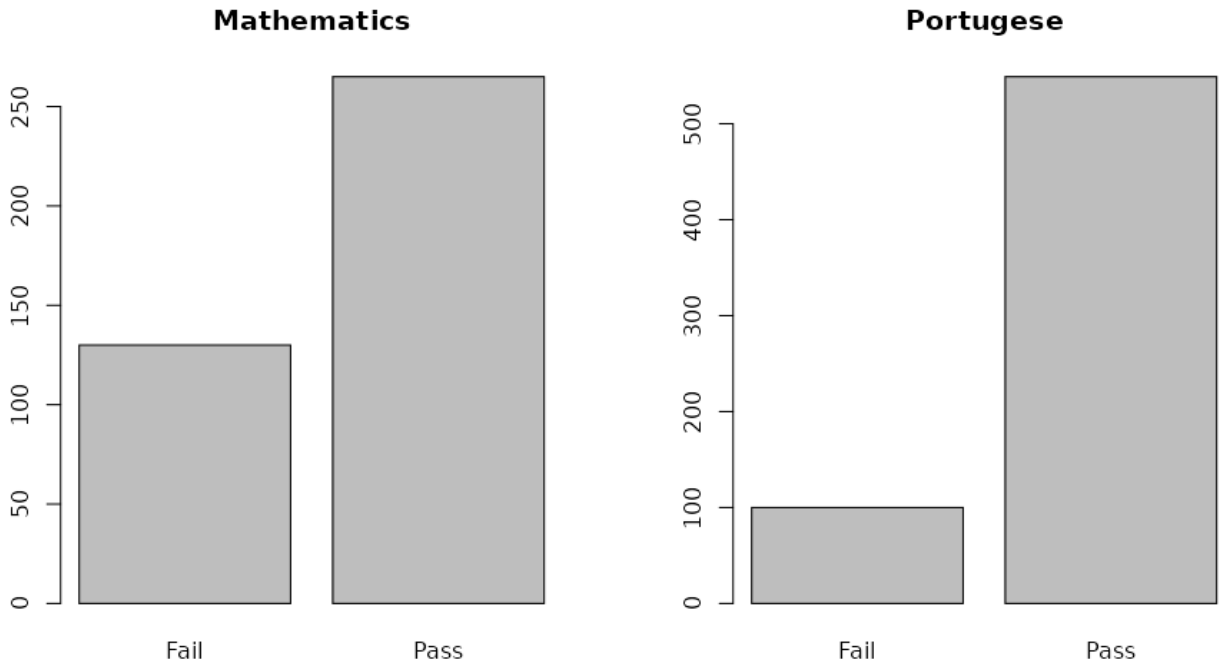
	Fail	Pass	Accuracy	.9949
Fail	129	1	Sensitivity	.9923
Pass	1	264	Specificity	.9962

SVM Radial Por w All (Thres=0.190)

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

Overall, we can see how mathematics is the more efficient case to work on the prediction when compared to portugeuse. On the mathematics side, the performance of both SVM kernel models are significant, especially the Radial kernel. Same can be done to SVM Radial kernel performance with portuguese, but it might be an overfitted model. We'll never know until actually applying the model to the real world test data. Like it was mentioned in the SVM Radial section, re-tuning of kernel parameters will probably help if the models were found to be overfitted when applied against real world data.

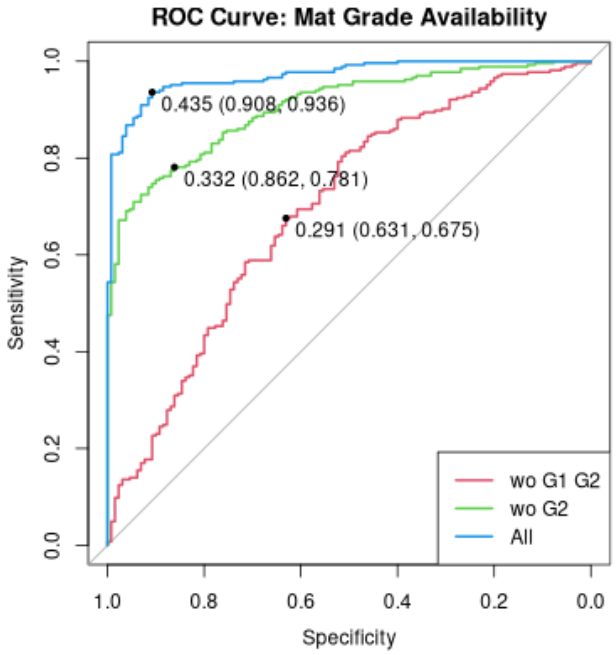
Thoughts on Dataset Sample Distribution ratio:

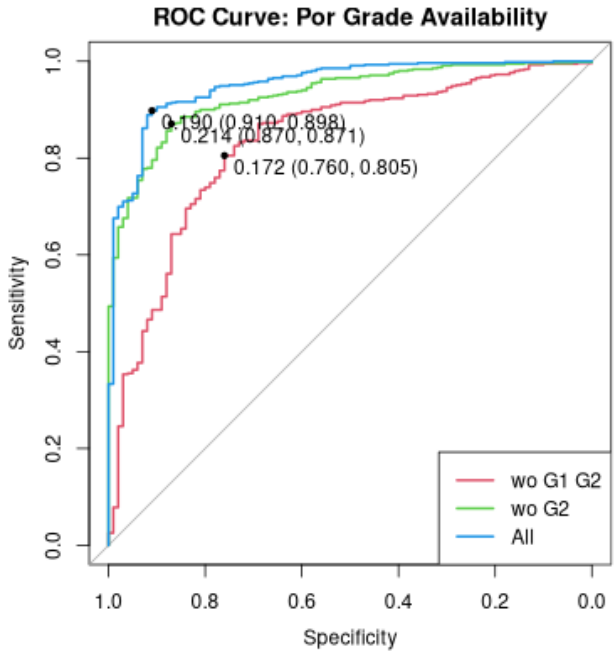


The bar plots above show the imbalanced sample distribution nature between two classes within each dataset. In comparison between the two datasets, the difference can be observed in the ratio of class distribution. The portuguese dataset has a very low ratio of fail samples in comparison to the mathematics dataset. Roughly, the Fail/Pass ratio in Mathematics is about 1:2, and about 1:5 for portuguese. This imbalance must be one of the major reasons why prediction on portuguese seems more difficult compared to mathematics.

Thoughts on Grade Availability:

As a real world approach of the trained models from this search, grade availability differs depending on the time of the year the prediction is made. Since grade predictors G1 and G2 are strongly correlated to the prediction results based on the p-value, the prediction is expected to degrade when one or two of the grades is not available. As a bit of bonus research, we decided to look into how the grade availability affects the prediction with the best prediction rate model, SVM Radial kernel.





Above are the ROC curves of best trained SVM Radial kernel models, with variation in grade availability for model training and prediction for both datasets. Below are the confusion matrices with the above plotted models after optimal threshold application.

Confusion Matrices: Grade Availability

Mat w/o G1 G2

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

Por w/o G1 G2

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

Mat w/o G2

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

Por w/o G2

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

Mat w/ All

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

Por w/ All

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

The outcomes are somewhat confusing. Without any grade availability, prediction on Mathematics will not work. Poor performance of it can be seen from its ROC curve and prediction accuracy of 74%. In contrast, the model for portuguese is in somewhat of an usable shape which can be seen from its ROC curve. The accuracy of the prediction is at a surprising 98%. When the G1 grade became available, both ROC curve shape and prediction accuracy took a leap on Mathematics model performance, with prediction accuracy of 98%. On the other hand, portuguese model's ROC curve shows good improvement with its shape, but prediction accuracy goes down to 93%. This accuracy rate is not bad at all, but the fact that the prediction accuracy decreases from 98% to 93% does not make sense. It implies those samples which were predicted correctly, are not predicted incorrectly after G1 grade availability. When both G1 and G2 grades become available, the ROC curves and predictions accuracy improves further to the best SVM Radial kernel model we reached from this research. This had a prediction rate of 99% on both datasets.

Conclusion:

Using the SVM Radial Kernel Model, it proved to be the most effective in providing a high prediction of accuracy on both datasets. Possible causes that may account for the difference of general prediction accuracy rate between the datasets are that the dataset size in portuguese was almost double the size of the mathematics dataset. In addition, the Fail/Pass class sample distribution ratio within each dataset was found to be imbalanced, especially on the portuguese dataset. The difference of the ratio between the datasets is most likely due to the higher drop-out ratio in the mathematics dataset compared to the portuguese dataset. Imbalance of class distribution ratio, and sample size difference 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 the prediction model of portuguese was still found to be feasible without any grade availability. In retrospect, RMSE measurement was found incomparable when the predictor number or dataset used for model trains are different sets. For the model which favors a smaller number of predictors, application of few selected predictors were found effective. For instance, results from the KNN regression model displayed very poor performance with all predictors in the scatter plot. This confirms the nature of KNN does not work well with high predictor dimensions. 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 a 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 practiced the most favourable technique 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