Phuong Tang
ID: 23186540

Assignment 3- DATA423-20S1

1. Strategies of finding the best model

In this assignment, we employed the way of choosing candidate models from a variety of styles. To derive the knowledge about the good performance of regression methods, we read papers as a trusted reference source for us to choose a good subset of models from each style. After having an overview about the performance of each style by training a small sample of models, we defined which style is the most suitable style for our dataset. Then, we focused to dig down many methods from that style to find the optimal solution. That is our overall strategies of finding the best model.

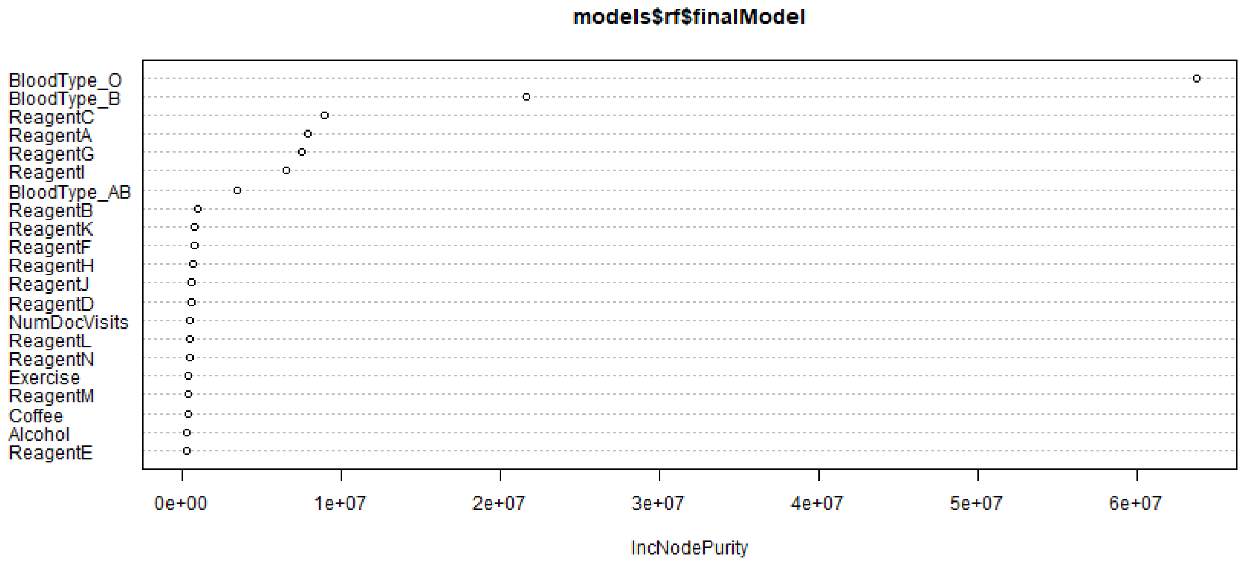
In our research, on the way of trying 3 kernel method, we figured out that Radial basis function is the most suitable style which can give us low RMSE. Therefore, we decided to concentrate on exploring as many methods related to Radial kernel as possible. The performance of Polynomial function is also quite good which can be tried along the way.

The list of methods which we trained in our assignment for each style illustrated as below:

- Kernel methods: subdivided into 3 styles:
    - Linear basis function: Kernel PLS, PLS, GLM Boost, GLMnet, elastic net
    - Polynomial basis function: krlsPoly (Polynomial Kernel Regularized Least Squares)
    - Radial basis function: krlsRadial (Radial Basis Function Kernel Regularized Least Squares), gaussprRadial (Gaussian Process with Radial Basis Function Kernel)
- Tree based: xgbTree, Rpart
- Neural network: nnet, rbf (Radial Basis Function Network)
- Random forest: rf (random forest), RRFGlobal (Regularized Random Forest)
- Support vector machine: svmRadial (Support Vector Machines with Radial Basis Function Kernel), svmPoly (Support Vector Machines with Polynomial Kernel)
- Relevance vector machine: rvmRadial

The strategies should come with a knowledge about the data itself. To have a quick glance about the dataset, random forest with VarImPlot can help us to have an overview about the variable importance evaluation. Looking at the figure 1, we can see top 5 useful variables which help to achieve higher increases in node purities are “Bloodtype_O”,” Bloodtype_B”, “ReagentC”, “ReagentA”, “ReagentG”. In other words, these are significant predictors which contribute to predict Y correctly in our regression model.

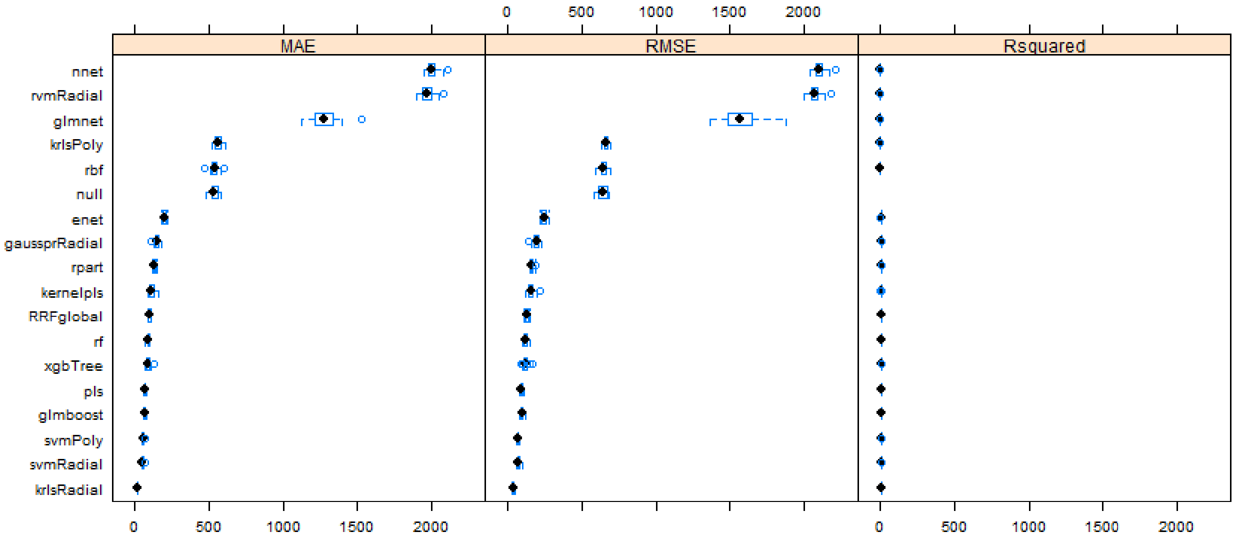
Figure 1: The variable importance plot retrieved from Random Forest method.



2. Results

The summary of model selection as shown in figure 2. There are 17 models have been trained and compared based on 3 key metrics RMSE, MAE and R-squared. Overall, krlsRadial, svmRadial and svmPoly are top 3 models have best performance in our regression problem. Out of these, krlsRadial (Radial Basis Function Kernel Regularized Least Squares) is the best performing model with minimum RMSE only 34.83.

Figure 2: Model selection visualization in assignment 3



The summarization about the methods, preprocessing steps employed and resampling results including final hyperparameters and metrics are illustrated in table 1.

Table 1: Summarized table about methods, preprocessing steps employed and resampling results.

No	Method	Pre-processing step	Resampling results			
			Hyperparameter	**RMSE**	**MAE**	**R-squared**
1	NULL	none	none	636.85	537.67	NA
2	GLMnet	naomit, dummy	Alpha=0.87 Lambda = 5.72	1570.02	1278.15	0.02
3	PLS model	Knnimpute, dummy	Ncomp = 21	93.77	69.3	0.98
4	Rpart	none	Cp = 0.00	164.08	128.93	0.93
5	Kernel PLS	Knnimpute, dummy	Ncomp = 19	156.31	115.4	0.94
6	GLMboost	Knnimpute, nzv, dummy	Mstop = 403 AIC prune = yes	94.76	68.55	0.98
7	krlsRadial	Knnimpute, dummy	Lambda = 0 Sigma = 832.44	34.83	18.5	1
8	krlsPoly	knnimpute, dummy, center, scale	Lambda = 0.07 Degree = 2	660.99	562.72	0.09
9	svmRadial	Knnimpute, dummy	Sigma= 0.01 C = 90.66	73.18	52.81	0.99
10	svmPoly	Knnimpute, dummy, center, scale	Degree = 3 Scale = 0.01 C = 14.12	68.06	55.78	0.99
11	rvmRadial	Knnimpute, dummy	Sigma = 0.01	2074.67	1975	0.01
12	gaussprRadial	Knnimpute, dummy	Sigma = 0.02	192	150.15	0.94
13	nnet (Neural network)	Knnimput, center, scale, nzv	Size = 2 Decay = 0.1	2103.25	2006.42	0.04
14	rbf (Radial basis function Network)	Knnimpute, dummy	Size = 13	643.63	540.05	0.01
15	rf (random forest)	Knnimpute, dummy	mtry = 12	119.94	89.39	0.97
16	RRFGlobal (Regularized Random Forest)	Knnimpute, dummy	mtry =14 coefReg = 0.3	126.82	96.74	0.96
17	xgbTree	Knnimpute, dummy	eta=0.1, max_depth=25 gamma=0, colsample_bytree=0.9 min_child_weight=1 subsample =1 nrounds = 200	118.61	89.08	0.97
18	enet (elastic net)	Knnimpute, dummy, center, scale	Lambda = 0.03 Fraction = 0.56	245	201.11	0.92

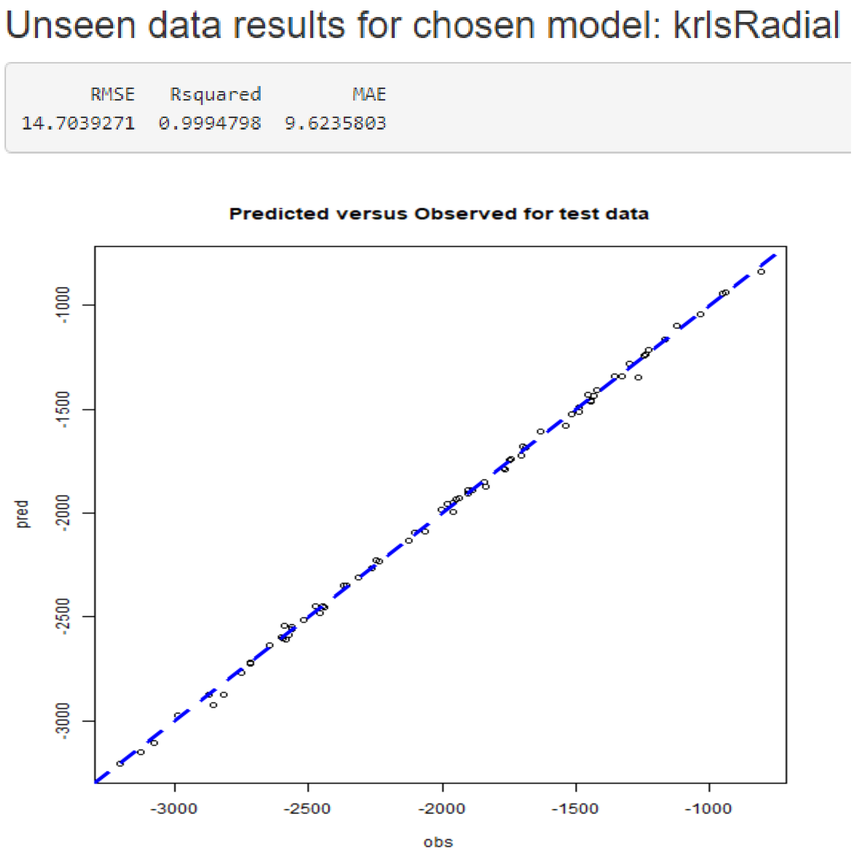
3. Dig down about the best model

Let’s try to figure out why krlsRadial model fit our data so well. The first impression is when we fit our data by using GLM model, the result is not good as opposed to Radial basis kernel function. In general, GLM usually impose a strict functional form assumption to the dataset while KRLS (Kernel Regularized Least Square) approach offers a versatile modelling tool that strike to compromise between a highly constraint GLM and a more flexible but often less interpretable machine learning approaches by obtaining a continuously differentiable solution surface. GLM assumes that the outcome is a weighted sum of the independent variables. In contrast, KRLS is based on the premise that information is encoded in the similarity between observations, with more similar observations expected to have more similar outcomes (Hainmueller and Hazlett, 2014). Our data set is multidimensional with 20 variables. Looking at the PLS model, we need more than 16 principal components to reconstruct the data, this means that our data are highly uncorrelated. Therefore, a flexible regression model such as krlsRadial or PLS will fit the model better than strict functional form such as GLMnet.

However, why only Radial basis kernel (Gaussian kernel) fit data well while polynomial kernel doesn’t. Gaussian and Polynomial kernel are simply different in case of making the hyperplane decision boundary shape. This means our dataset fit with a “bell curve” shape than polynomial shape.

Figure 3 shows the performance of krlsRadial model on test data. The model fit on test data very well with low RMSE and high R-squared.

Figure 3: The performance of krlsRadial model on test data



Unseen data results for chosen model: krlsRadial

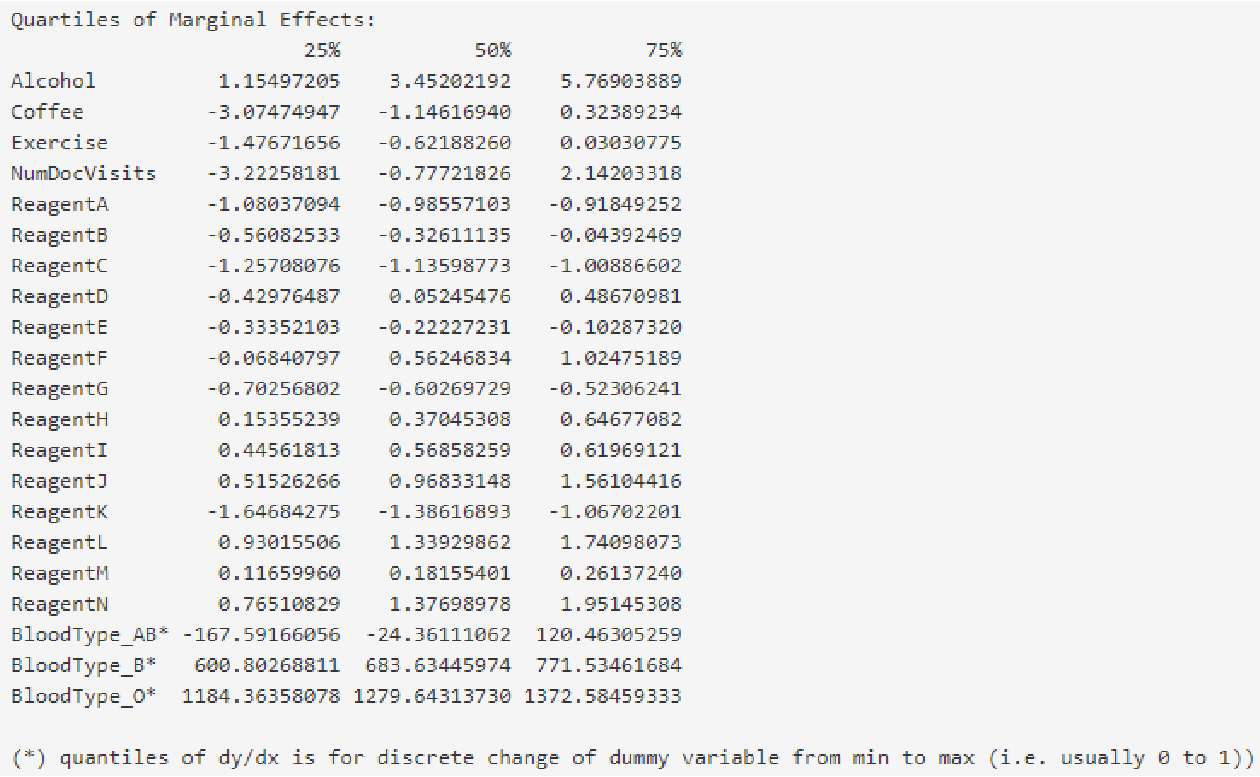
```
      RMSE   Rsquared        MAE
14.7039271  0.9994798  9.6235803
```

4. How would the optimum model change if transparency were very important?

Based on (Hainmueller and Hazlett, 2014), krlsRadial model allows interpretation in ways analogous to generalized linear models. KRLS analyses provides an estimate of the average pointwise marginal effect (like β coefficient from linear regression and could be interpreted as the average marginal effect) for each independent variable along with heterogeneity in the marginal effect expressed as an interquartile range (25th-75th). Therefore, the optimum model wouldn't change as krlsRadial was still a transparent model.

For example, looking at the model summary of krlsRadial at figure 4, we can interpret that the marginal effect of "ReagentB", "ReagentD", "ReagentF", "ReagentH" are close to zero, which are accurate given that these predictors are indeed irrelevant for the outcome. On the other hand, the marginal effect of "BloodType" varies greatly in magnitude from the first quartile to the third quartile. This accurately captures the nonlinearity in the true effect of "BloodType" variable to the outcome.

Figure 4: Model summary of krlsRadial method



```
Quartiles of Marginal Effects:
                      25%          50%          75%
Alcohol         1.15497205   3.45202192   5.76903889
Coffee         -3.07474947  -1.14616940   0.32389234
Exercise       -1.47671656  -0.62188260   0.03030775
NumDocVisits   -3.22258181  -0.77721826   2.14203318
ReagentA       -1.08037094  -0.98557103  -0.91849252
ReagentB       -0.56082533  -0.32611135  -0.04392469
ReagentC       -1.25708076  -1.13598773  -1.00886602
ReagentD       -0.42976487   0.05245476   0.48670981
ReagentE       -0.33352103  -0.22227231  -0.10287320
ReagentF       -0.06840797   0.56246834   1.02475189
ReagentG       -0.70256802  -0.60269729  -0.52306241
ReagentH        0.15355239   0.37045308   0.64677082
ReagentI        0.44561813   0.56858259   0.61969121
ReagentJ        0.51526266   0.96833148   1.56104416
ReagentK       -1.64684275  -1.38616893  -1.06702201
ReagentL        0.93015506   1.33929862   1.74098073
ReagentM        0.11659960   0.18155401   0.26137240
ReagentN        0.76510829   1.37698978   1.95145308
BloodType_AB* -167.59166056  -24.36111062  120.46305259
BloodType_B*   600.80268811  683.63445974  771.53461684
BloodType_O*  1184.36358078 1279.64313730 1372.58459333

(*) quantiles of dy/dx is for discrete change of dummy variable from min to max (i.e. usually 0 to 1))
```

5. Recommendation about an ensemble of the top models

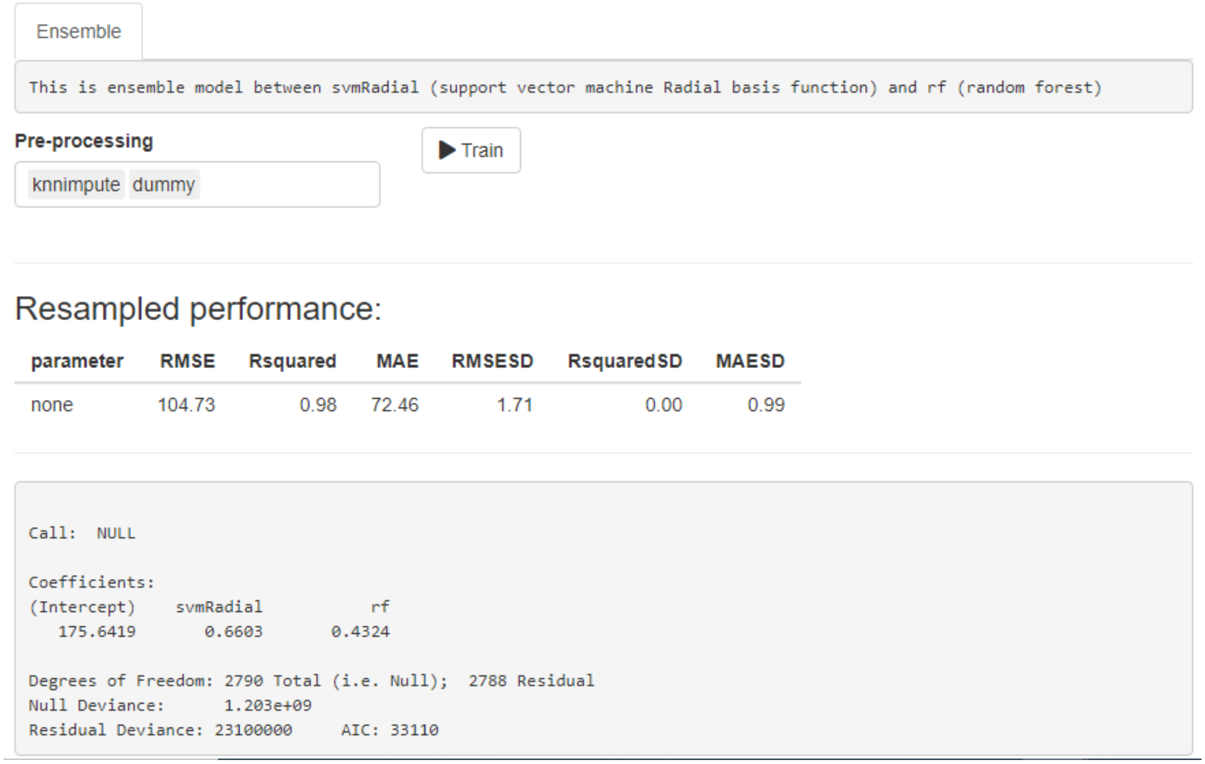
In our case, it is suggested that we shouldn't ensemble of top models in order to increase the accuracy of the regression model for the number of reasons listed as below:

- Ideally, we would ensemble models that are low correlated with each other. In our case, our top 3 models which are krlsRadial, svmRadial and svmPoly are correlated in a way of exploiting radial basis kernel and support vector machine algorithms.
- Reduction in model interpretation ability due to increased complexity and makes it very difficult to get any crucial data insight at the end.
- Computation time can be very high.

However, as a curiosity, with a spirit of “If you never try, you will never know”, we have tried to ensemble 2 models which are “svmRadial” and “rf” together using caretList and caretEnsemble. The RMSE is 104.73 which is not bad. However, this model is not the best performing model as compared to krlsRadial.

Quote: In R shiny, we would like to hide the coding for the tab “Ensemble” as it made errors in “Model selection” tab and “Performance” tab. The effort has been used to fix this problem but it ‘s not successful. I’m sorry T_T

Figure 5: Ensemble model tab



```
This is ensemble model between svmRadial (support vector machine Radial basis function) and rf (random forest)
```

Pre-processing

knnimpute dummy

Resampled performance:

parameter	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
none	104.73	0.98	72.46	1.71	0.00	0.99

```

Call:  NULL

Coefficients:
(Intercept)    svmRadial           rf  
   175.6419       0.6603       0.4324  

Degrees of Freedom: 2790 Total (i.e. Null);  2788 Residual
Null Deviance:      1.203e+09 
Residual Deviance: 23100000     AIC: 33110
```

References

a. Jens Hainmueller, Chad Hazlett. 2014. Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach https://web.stanford.edu/~jhain/Paper/PA2014a.pdf

b. Thomas J. Leeper. 2018. Interpreting Regression Results using Average Marginal Effects with R's margins. Retrieved from link: https://cran.r-project.org/web/packages/margins/vignettes/TechnicalDetails.pdf

c. Simple guide for ensemble learning methods – Juhi. Retried from link: https://towardsdatascience.com/simple-guide-for-ensemble-learning-methods-d87cc68705a2