Predicting the Quality of White Wine

Presented by:

Anil Silwal, PhD Student in Computer Science

Sumit Shrestha, Master's Student in Computer Science

Michigan Technological University

Layout

- Review of Datasets
- Data splitting and resampling
- Linear Models
- Non Linear Models
- Comparison and Analysis of Model

Datasets

UCI Machine Learning Repository
https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

White wine samples - 4898 observations of 12 variables.

Predictors

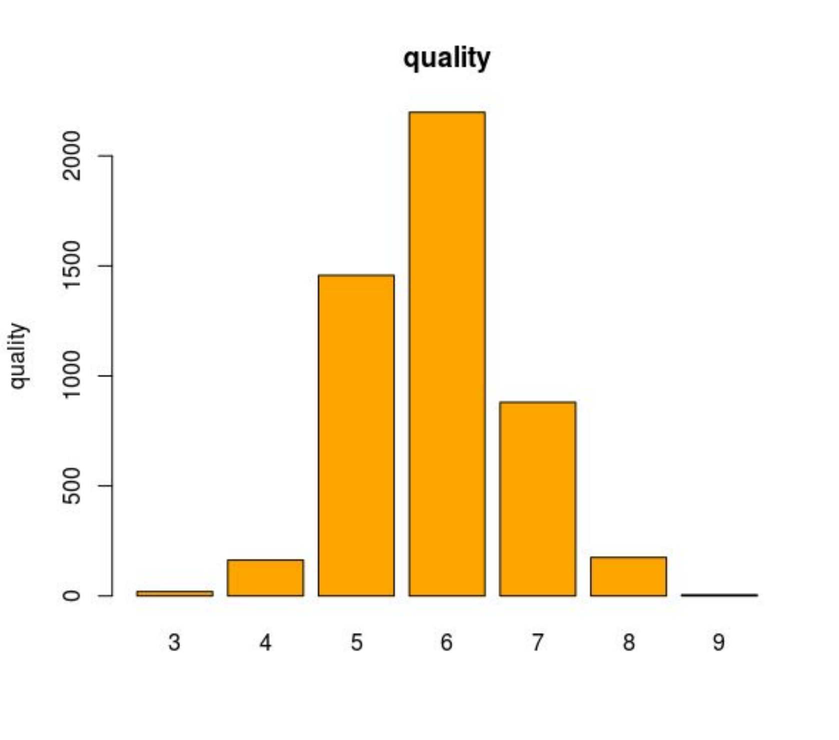
- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- pH
- Sulphates
- Alcohol - num(Data Type)

11 continuous predictor variables

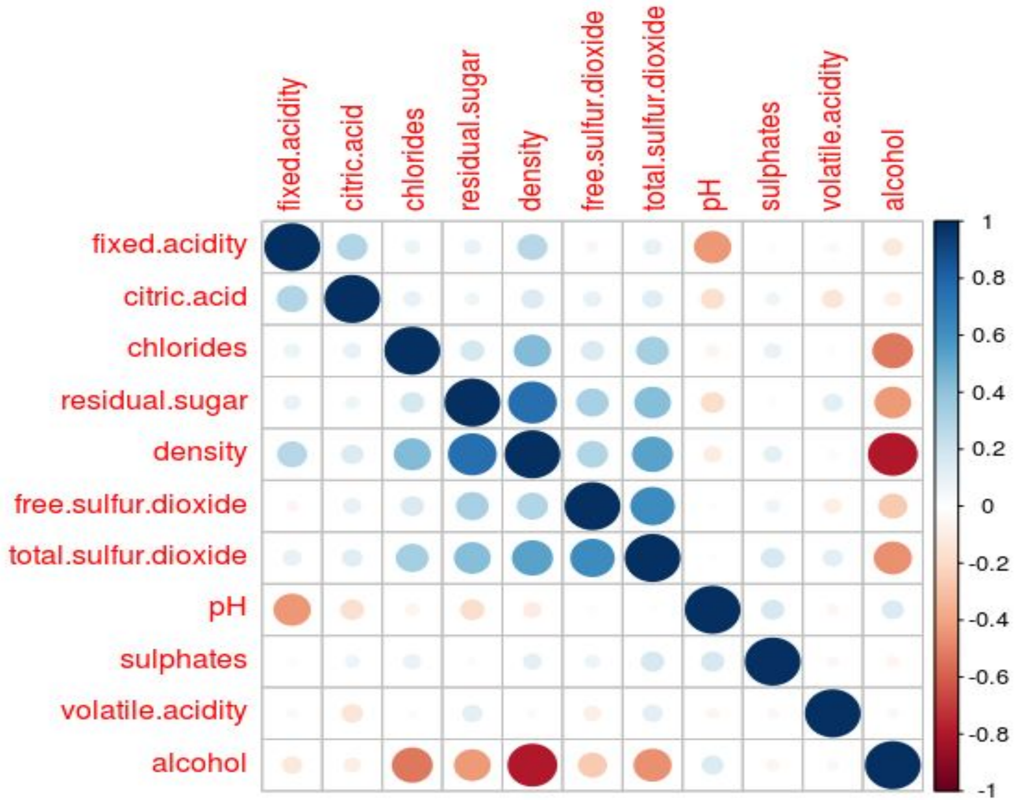
Response Variable

Quality - int variable (1 to 10)

Class	3	4	5	6	7	8	9
Count	20	163	1457	2198	880	175	5



Correlation Plot



Highly Correlated Predictors

Predictors	**Correlation value**
Density and residual sugar	0.7575
Alcohol and density	-0.7920

Correlation Analysis

<u>Not many highly correlated predictors</u>

- Density vs residual sugar (=0.75)
- Density and alcohol(=-0.79)

<u>Results of PCA</u>

9 principal components explained about 96% of variance
10 principal components explained about 99% of variance

Correlation Analysis

- 99% of the variation in data is shown by 10 PCs
- PCA not so feasible, since we have 11 predictors.
- Not much desired dimensionality reduction

Correlation cut off of 0.75 showed only one predictor to be removed

So entire data is not highly correlated. No any predictors removed.

Data Splitting

Stratified sampling preferred over random sampling
Training/Test data ratio = 80/20

Resampling:
Repeated K-fold cross validation with k=10 and repeats = 3 chosen

Building the model

- Linear regression model
- Non linear regression model

Due to the nature of our dataset both the problem could have also been done using the classification model

Linear regression model

- Ordinary Linear Regression
- Partial Least Squares
- Ridge Regression
- Elastic Net
- Lasso model

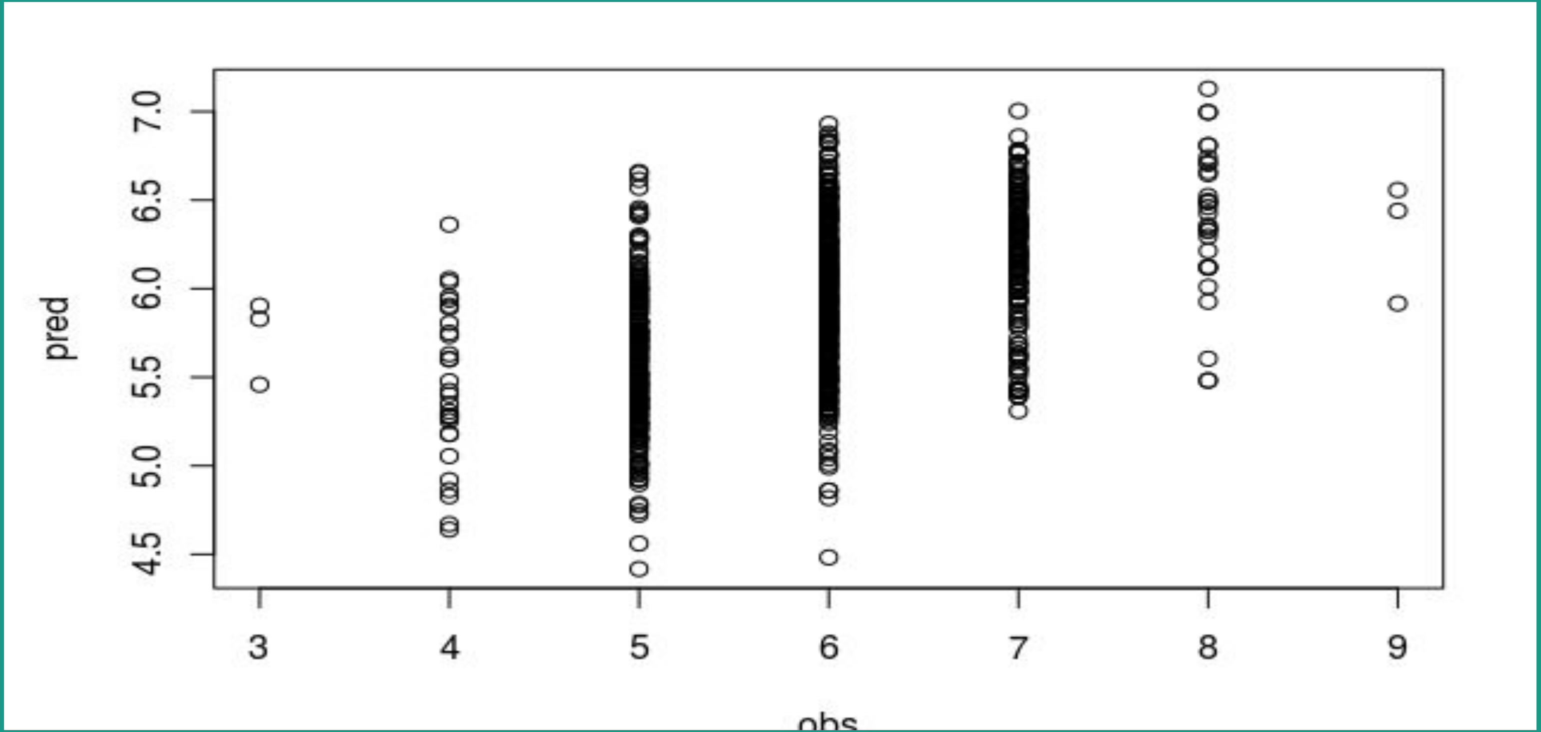
Ordinary Linear regression

PreProcessing:
Center, Scale and Box Cox Transformation

Resampling:
Cross Validation (10 fold repeated 3 times)

	RMSE	RSquared	MAE
Training Data	**0.7583111**	**0.2731315**	**0.5855021**
Testing Data	**0.7507707**	**0.2684493**	**0.5848243**

Ordinary Linear regression



Observed vs Predicted values for test data

Partial Least Square

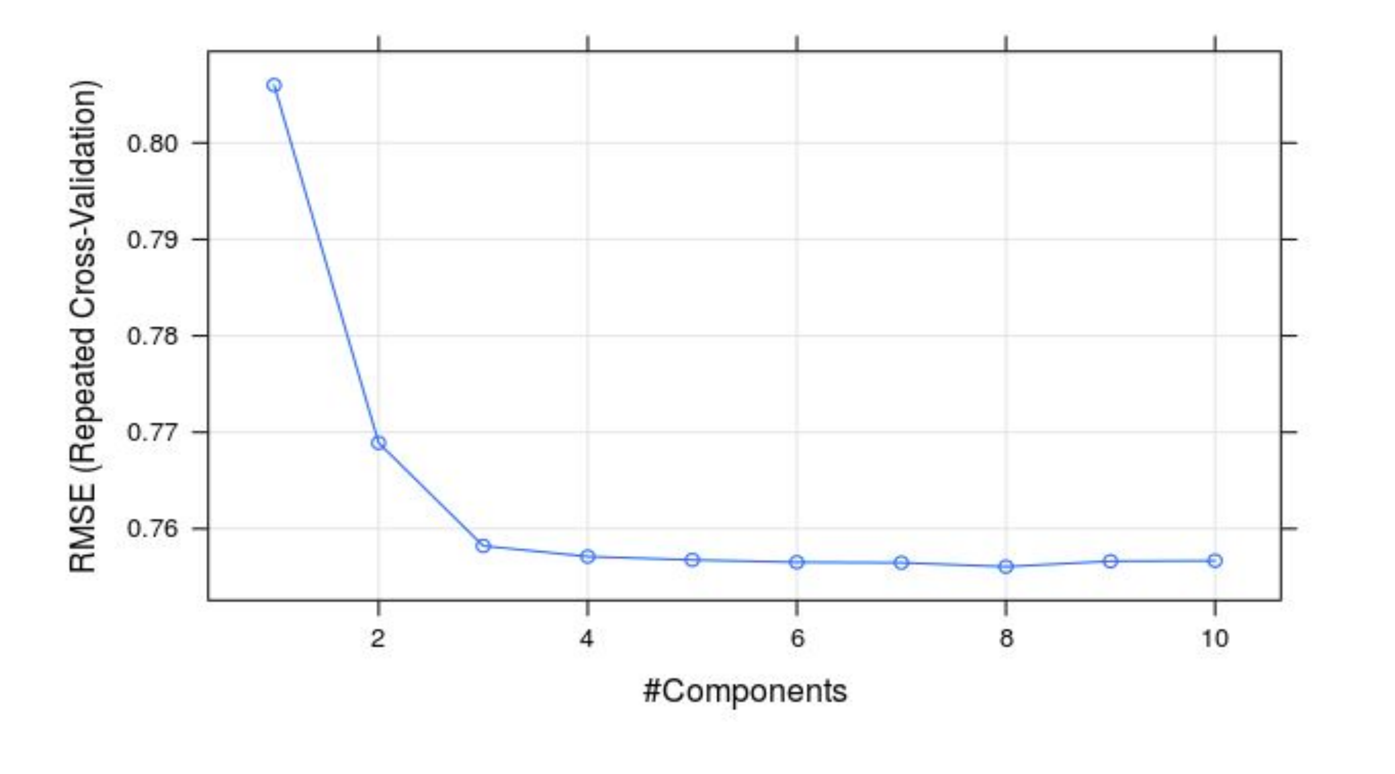
PreProcessing:
Center and Scale

Resampling:
Cross Validation (10 fold repeated 3 times)

Tuning Parameter: Number of components = 8

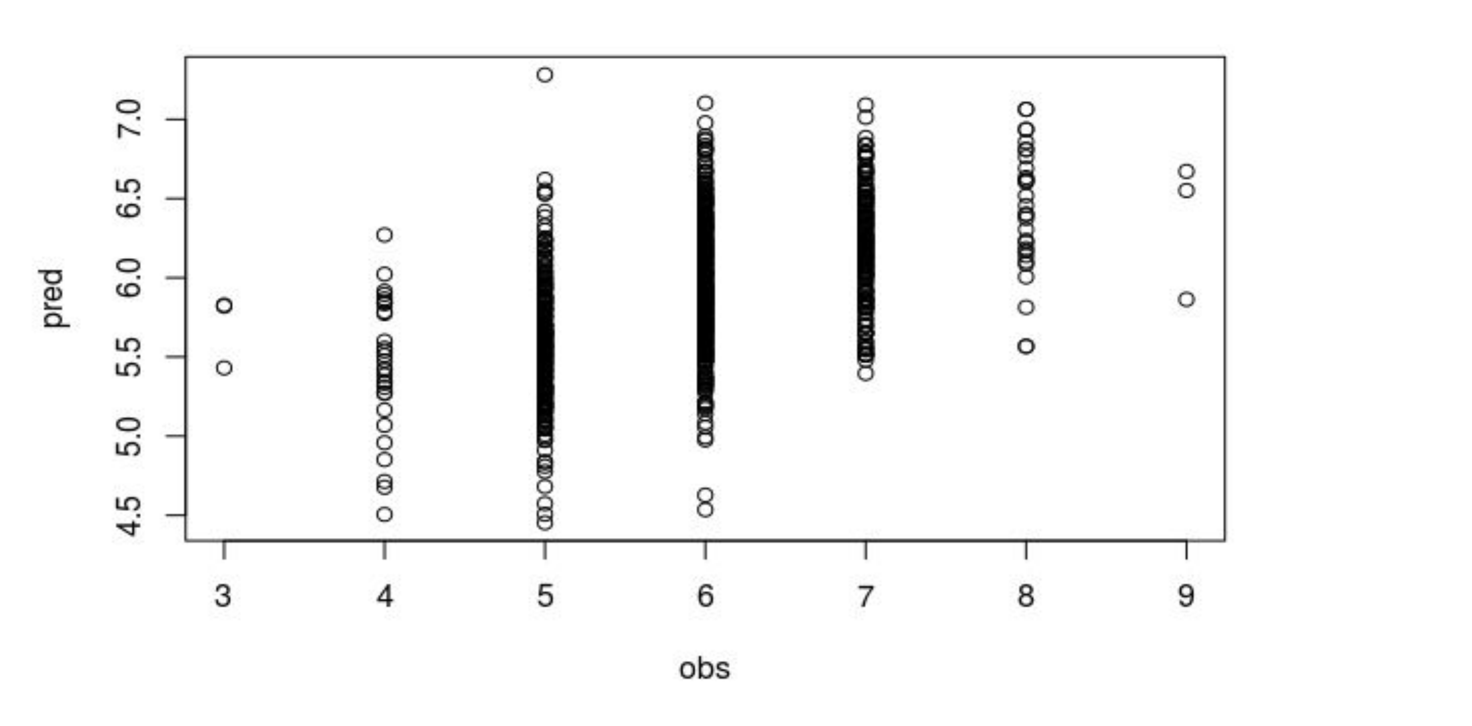
	RMSE	RSquared	MAE
Training Data	**0.7560663**	**0.2764547**	**0.5863016**
Testing Data	**0.7434357**	**0.2827261**	**0.5822419**

Partial Least Square



Number of PLS Components

Partial Least Square



Observed vs Predicted values on test data

Ridge Regression

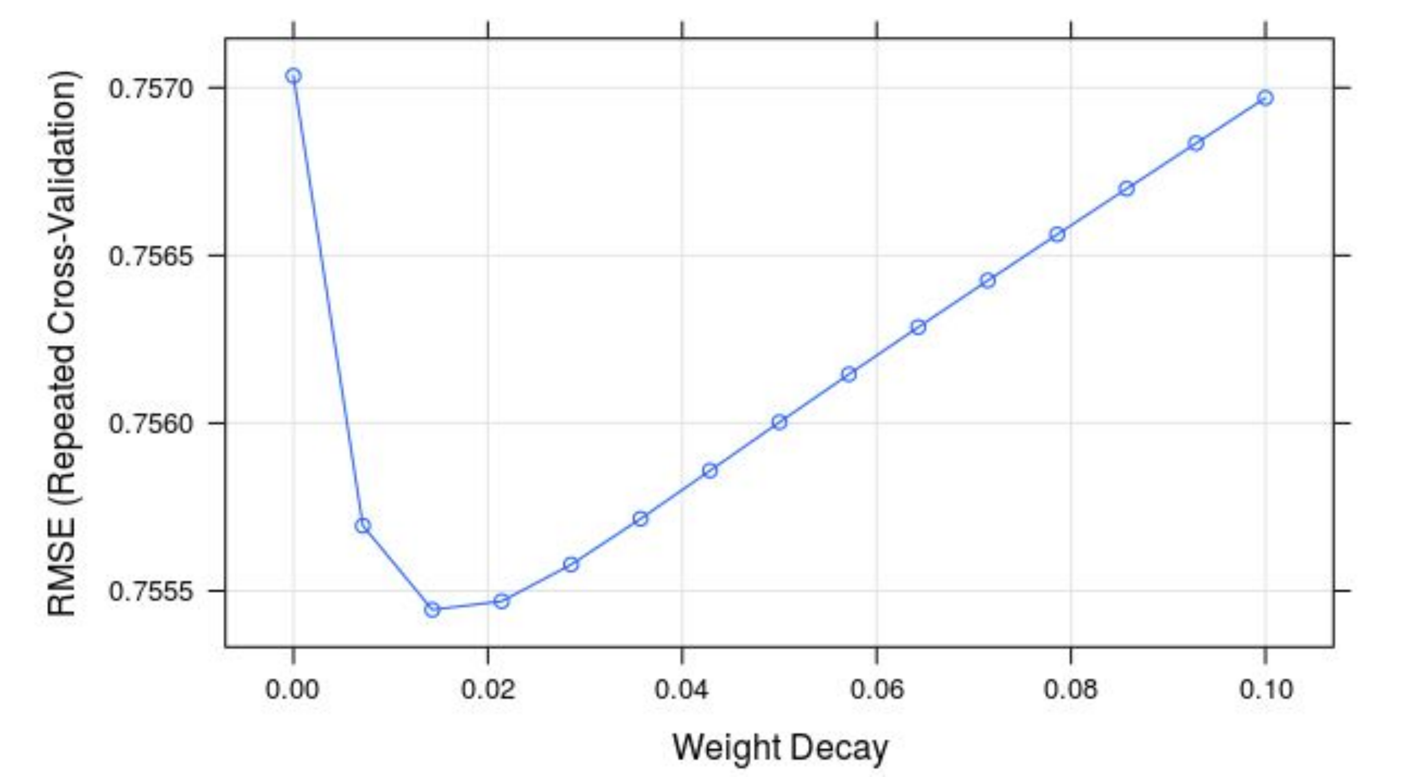
PreProcessing:
Centering and Scaling

Resampling:
Cross Validation (10 fold repeated 3 times)

Tuning Parameter: Lambda = 0.01428571

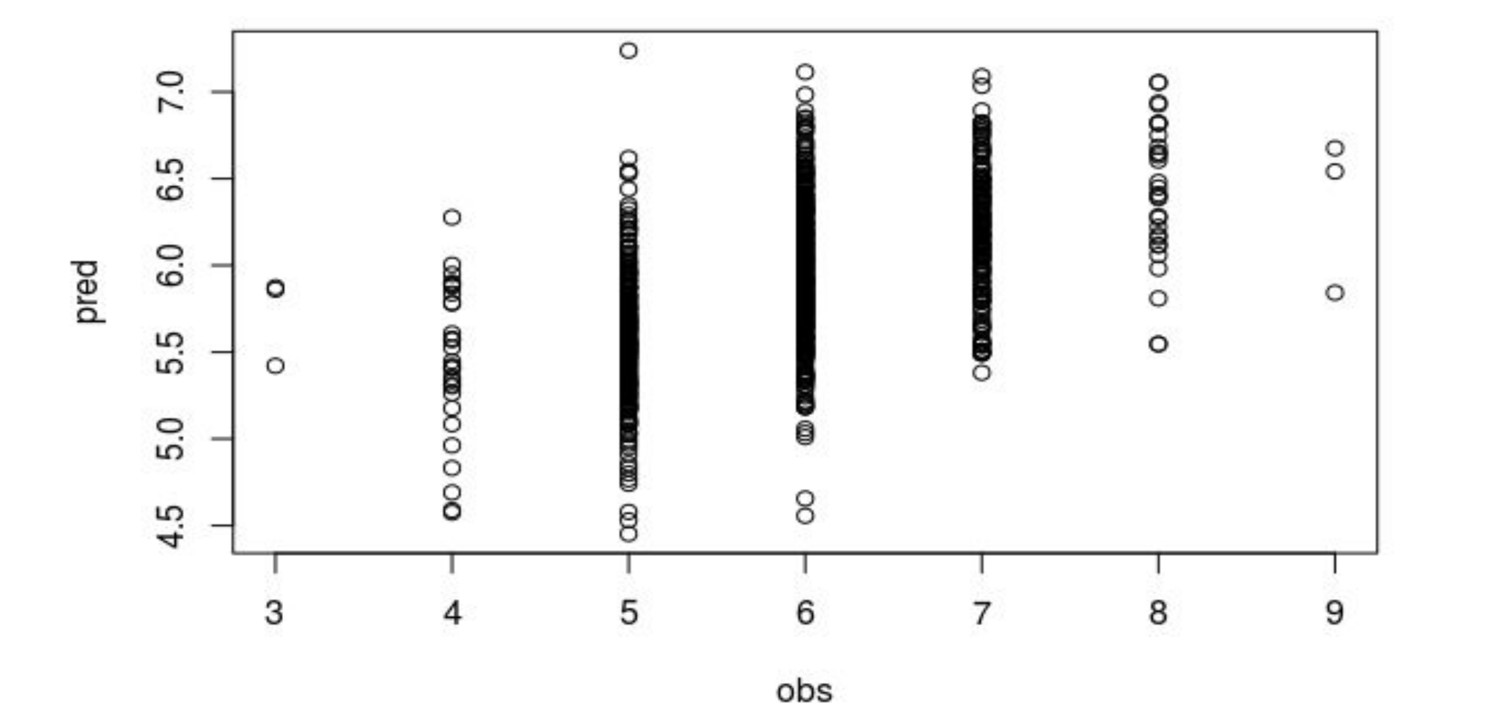
	RMSE	RSquared	MAE
Training Data	**0.7554437**	**0.2765135**	**0.5864161**
Testing Data	**0.7443214**	**0.2810610**	**0.5821881**

Ridge Regression



Different values of lambda

Ridge Regression



Observed vs Predicted values on test data

Elastic Net Model

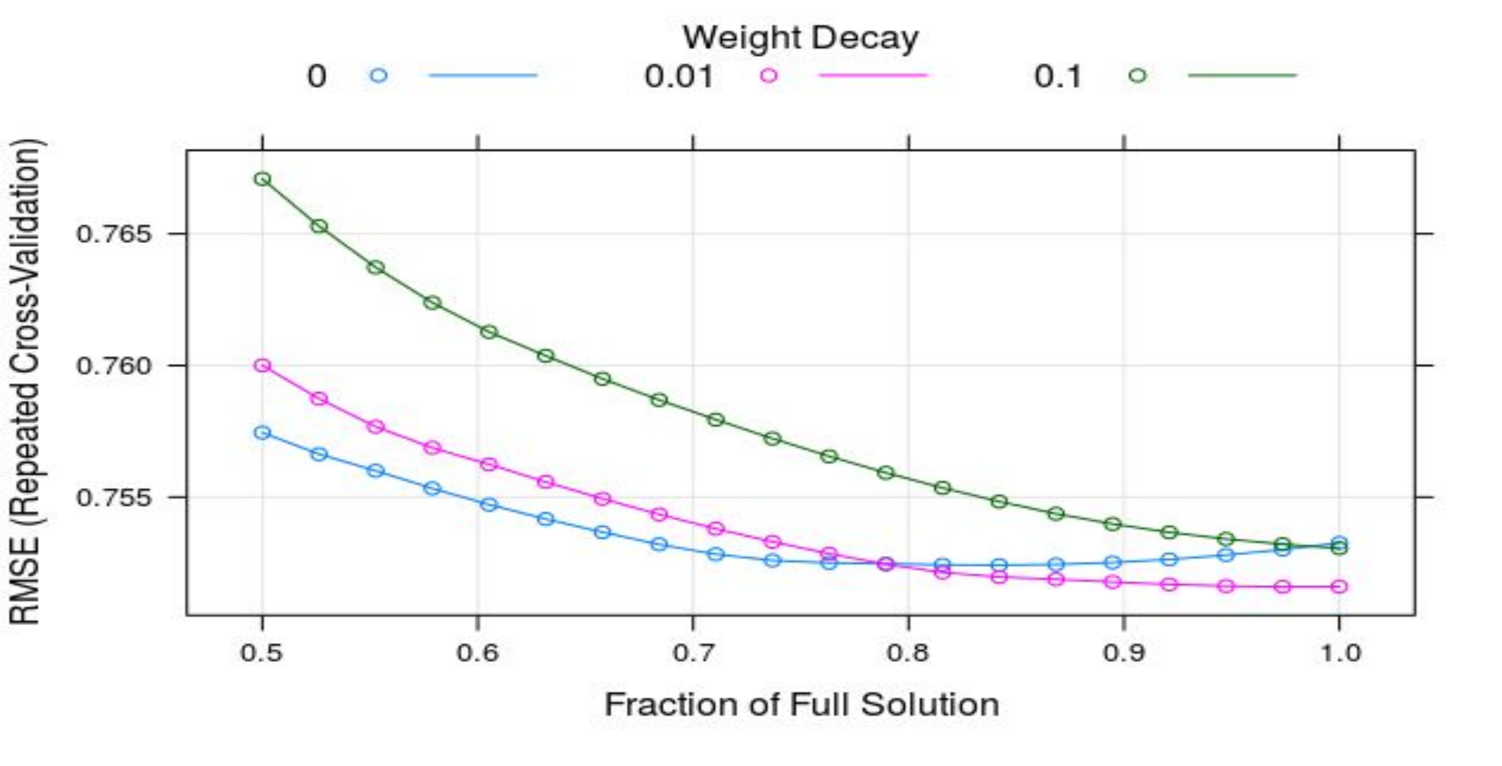
PreProcessing:
Centering and Scaling

Resampling:
Cross Validation (10 fold repeated 3 times)

Tuning Parameter: Fraction = 0.9736842 and lambda = 0.01

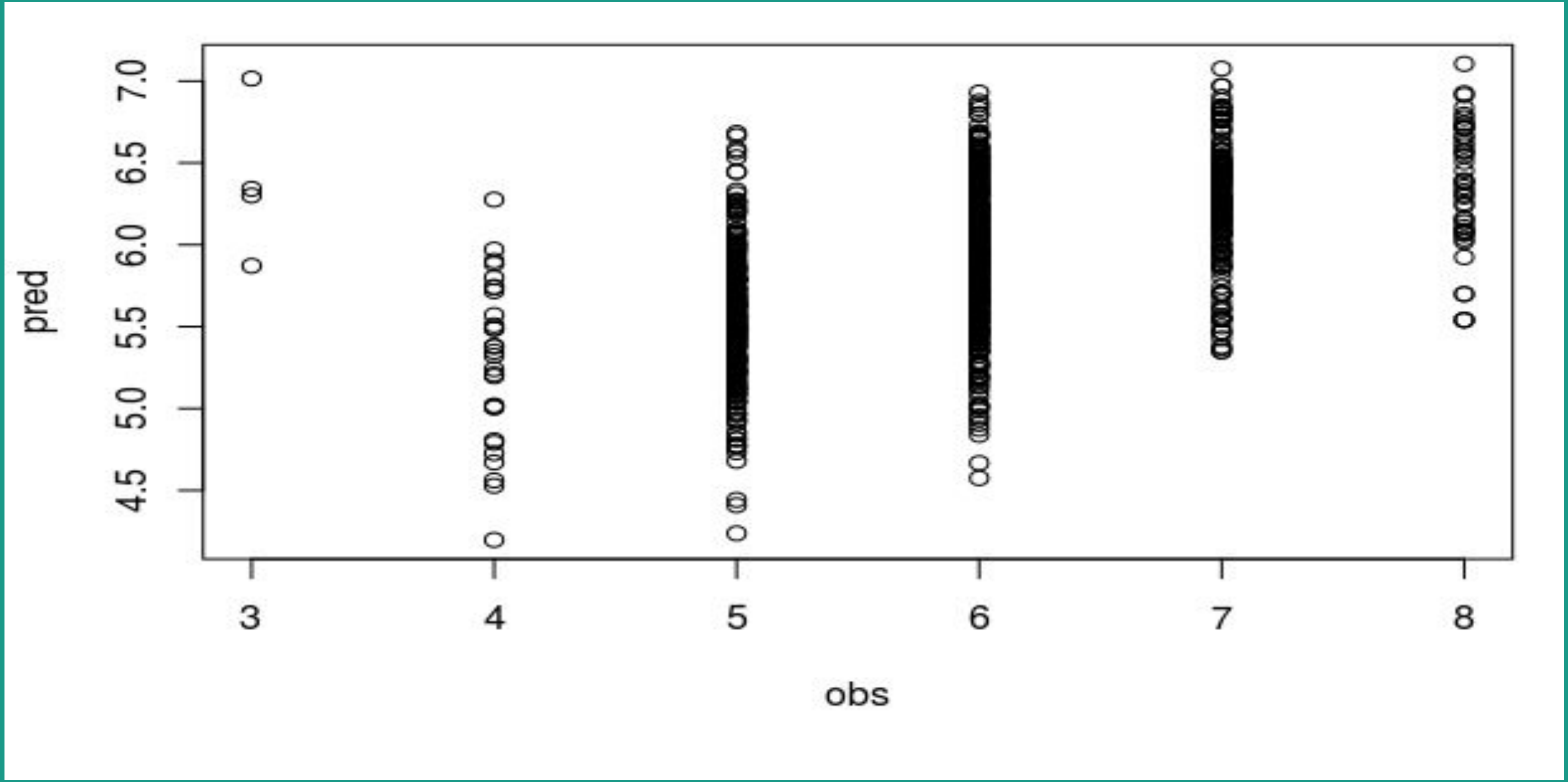
	RMSE	RSquared	MAE
Training Data	**0.7515978**	**0.2776304**	**0.5848679**
Testing Data	**0.7616966**	**0.2749985**	**0.5892220**

Elastic Net Model



Different values of lambda and fraction

Elastic Net Model



Observed vs Predicted values on test data

Lasso Model

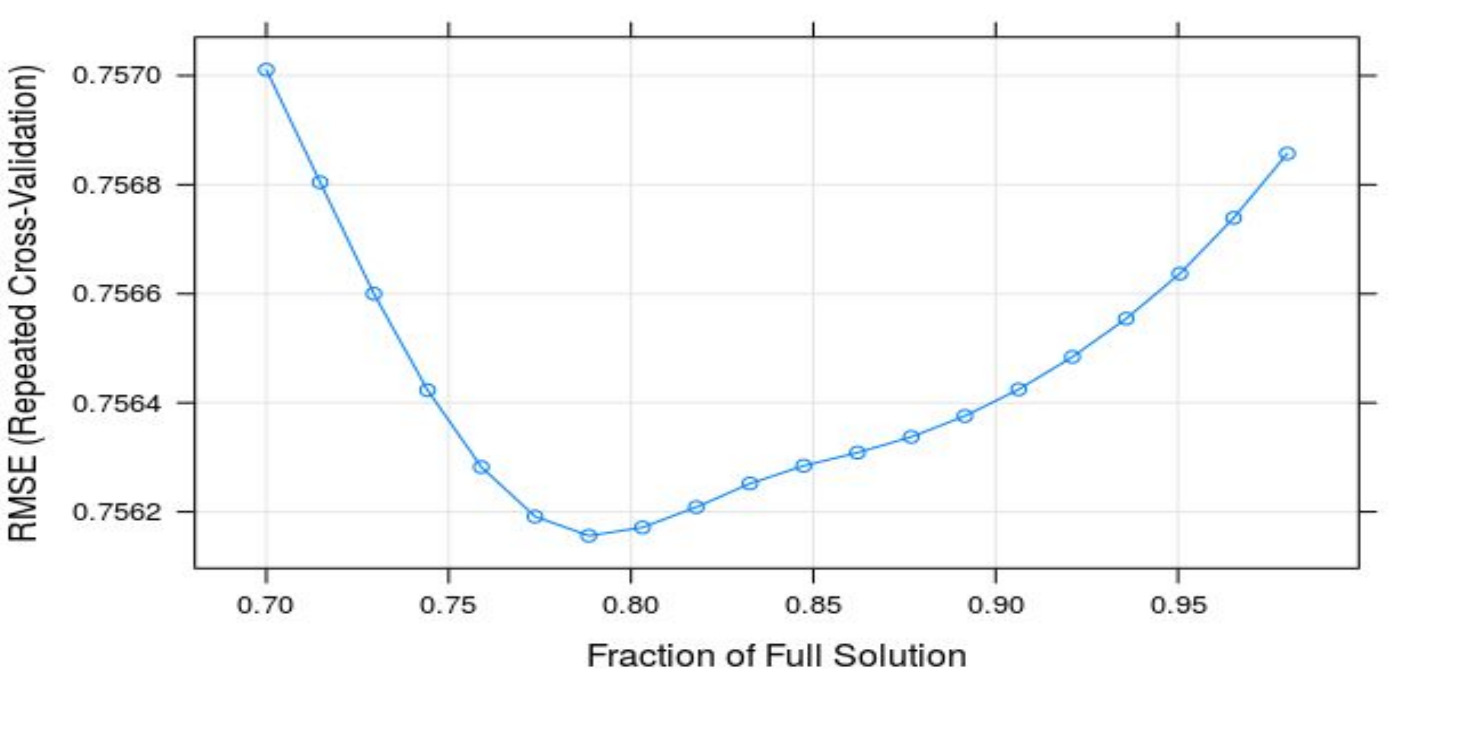
PreProcessing:
Centering and Scaling

Resampling:
Cross Validation (10 fold repeated 3 times)

Tuning Parameter: Fraction = 0.7884211

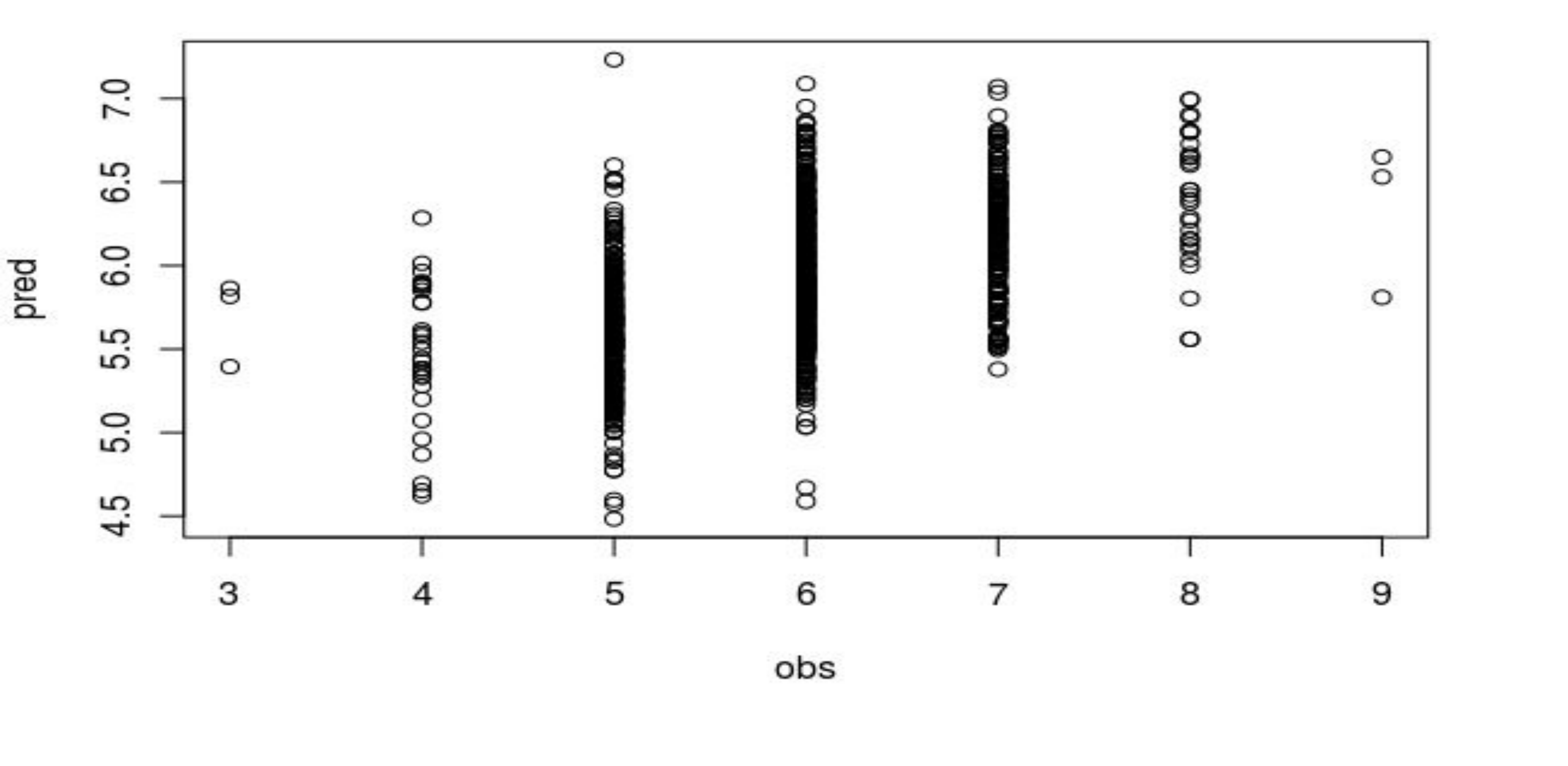
	RMSE	**RSquared**	**MAE**
Training Data	**0.7561563**	**0.2752103**	**0.5875141**
Testing Data	**0.7451390**	**0.2797990**	**0.5829684**

Lasso Model



Different values of fraction

Lasso Model



Observed vs Predicted values on test data

Non Linear regression model

- K Nearest Neighbours
- Neural Network
- MARS Model
- Support Vector Machine

K-Nearest Neighbour Model

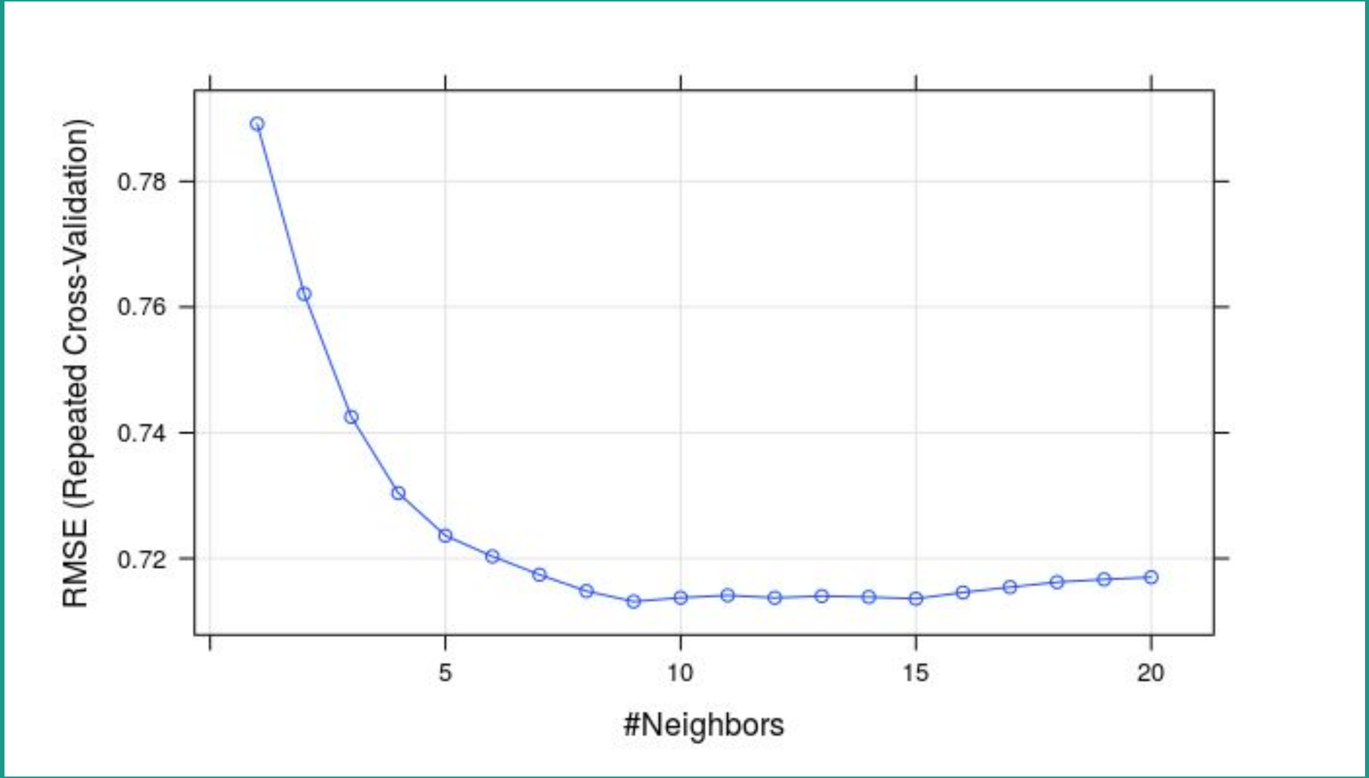
PreProcessing:
Centering and Scaling

Resampling:
Cross Validation (10 fold repeated 3 times)

Tuning Parameter: k = 9

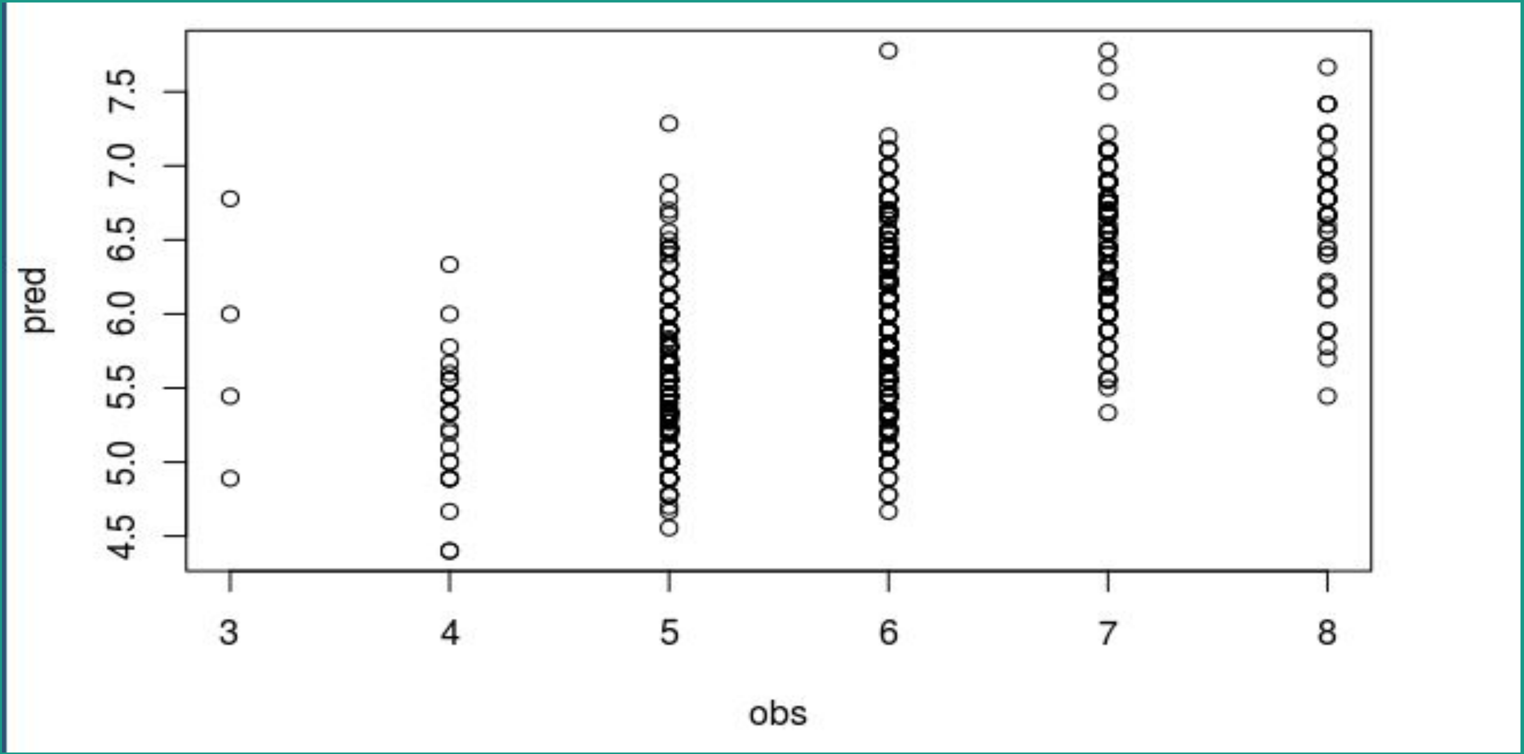
Training Data	RMSE	R^2	MAE
	0.7131459	0.3534372	0.5465718
Test Data	RMSE	R^2	MAE
	0.7092615	0.3741927	0.5411416

K-Nearest Neighbour Model



Tuning the number of neighbours

K-Nearest Neighbour Model



Observed vs Predicted values on test data

Neural Network Model

PreProcessing: Centering and Scaling
Tuning Parameter: size =9 and decay = 0.1

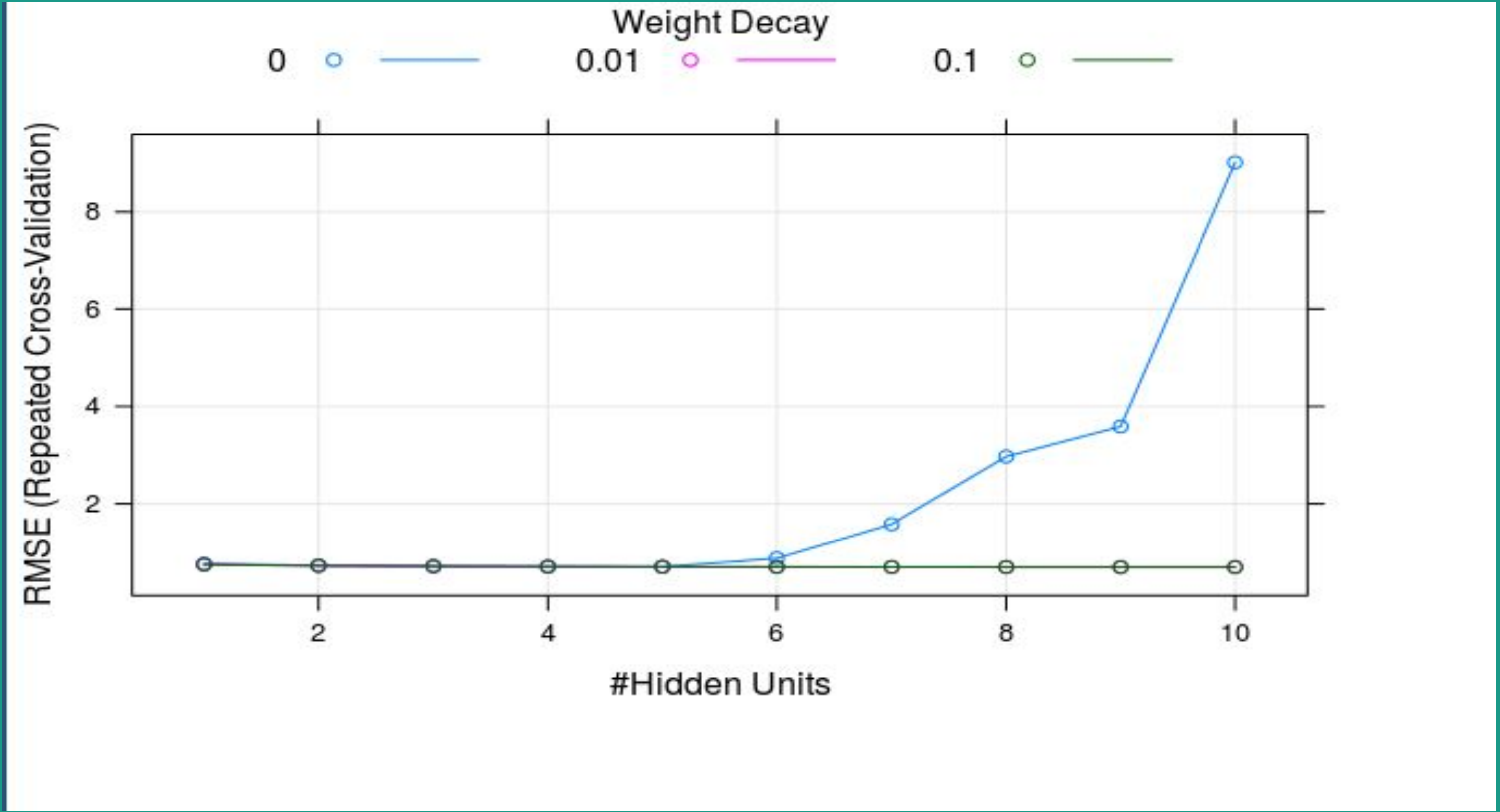
Training Data

RMSE	R^2	MAE
0.6949673	0.3822180	0.5416727

Testing Data

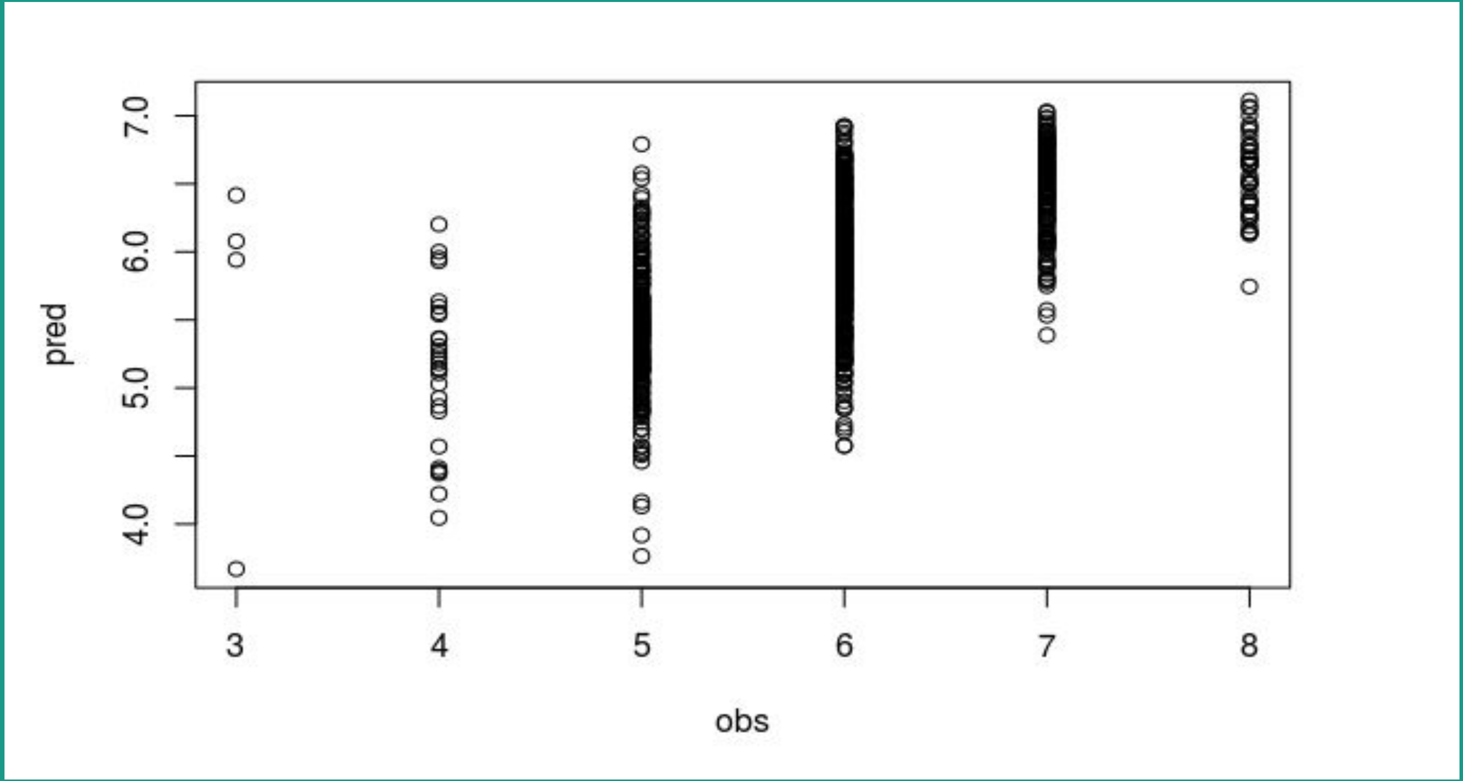
RMSE	R^2	MAE
0.6790271	0.4234195	0.5249017

Neural Network Model



Tuning the number of hidden units

Neural Network Model



Observed vs Predicted on the test data

MARS Model

Tuning Parameter: degree =2 and nPrune = 18

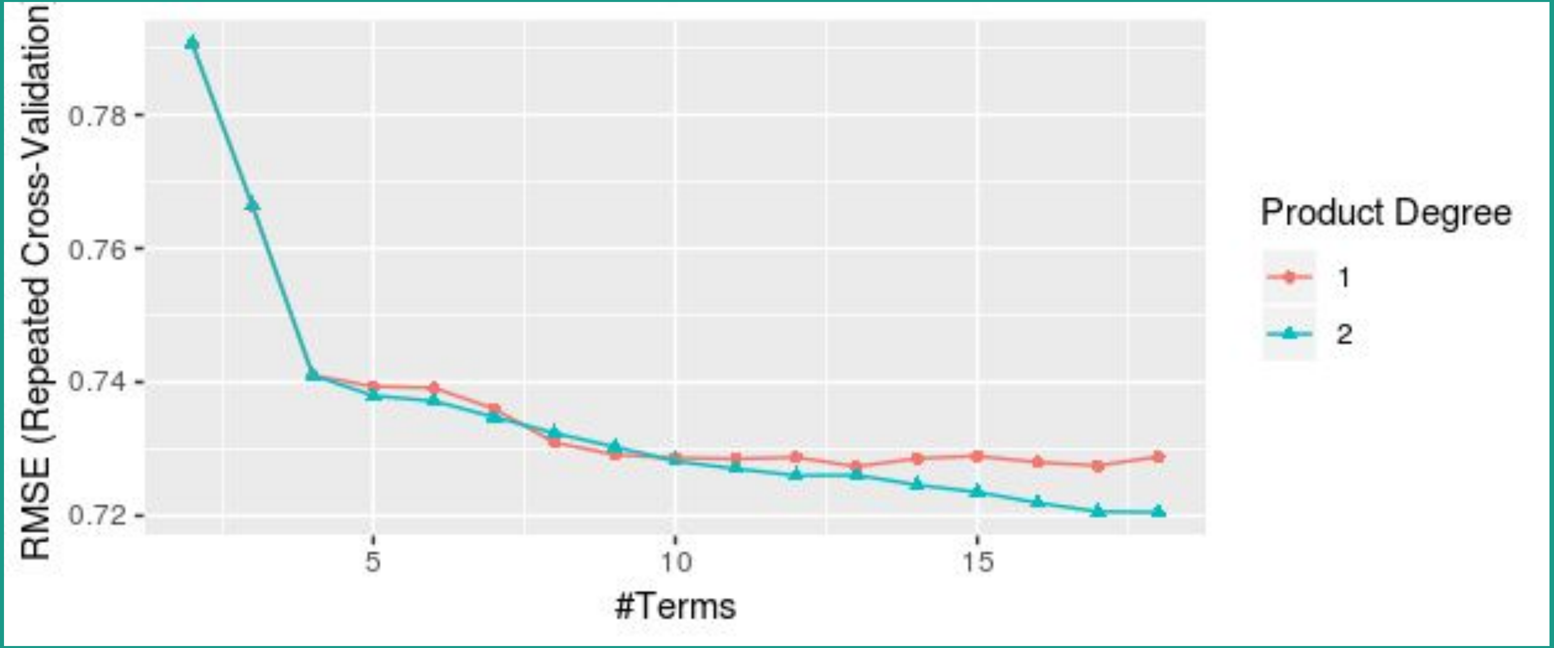
For Training Data

RMSE	R^2	MAE
0.7205307	0.3413517	0.5629384

For Testing Data

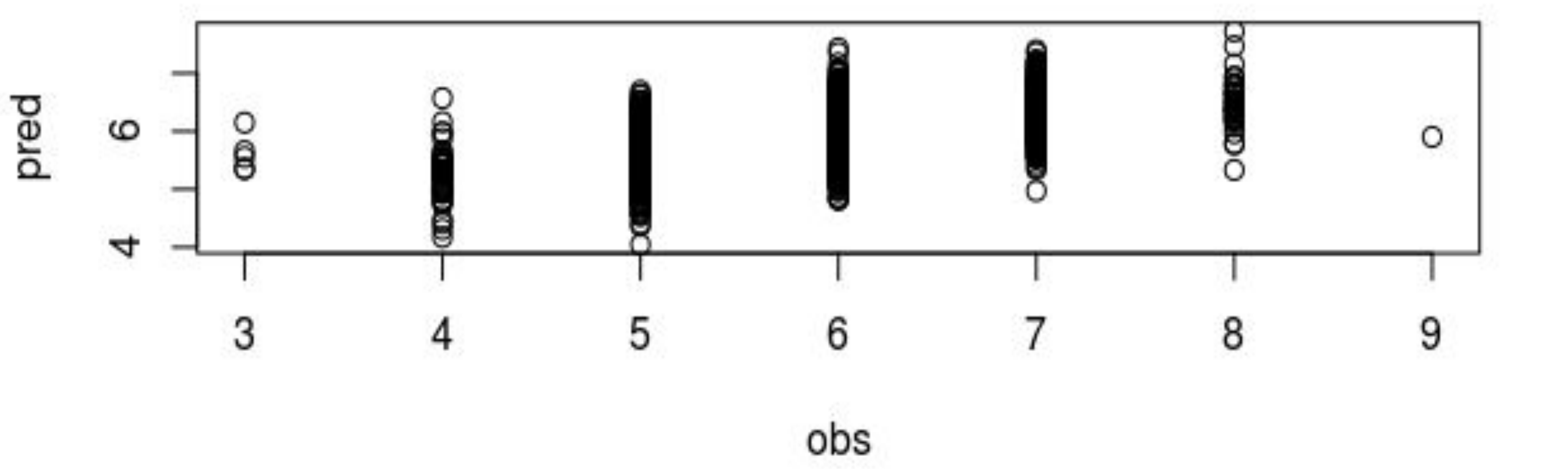
RMSE	R^2	MAE
0.7258565	0.3252741	0.5645746

MARS Model



Tuning the number of terms

MARS Model



Observed vs Predicted on the test data

Support Vector Machine

PreProcessing: Centering and Scaling
Tuning Parameter: sigma = 0.07934471 and Cost = 2

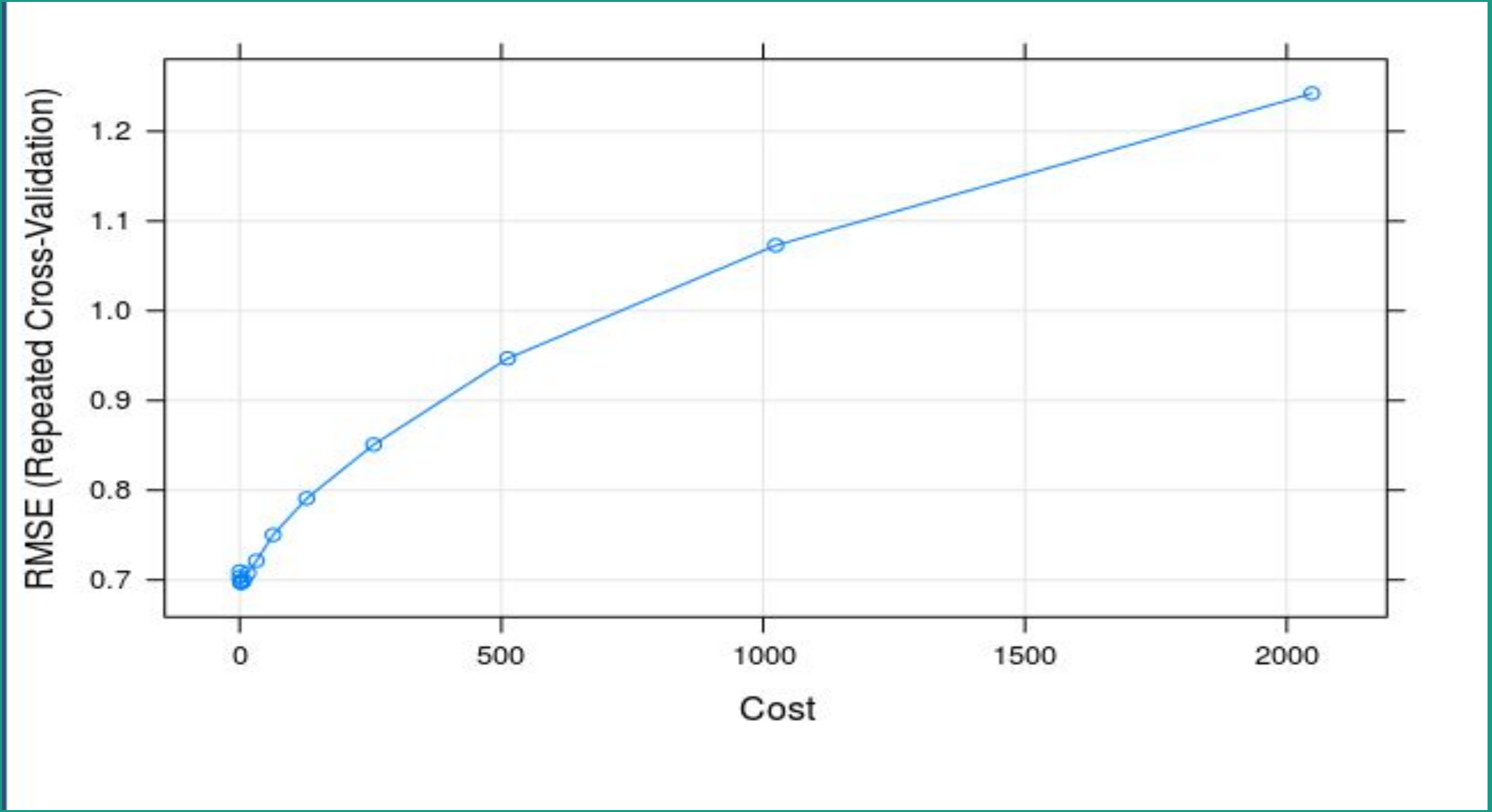
Training Data

RMSE	R^2	MAE
0.6966971	0.3816269	0.5241517

Testing Data

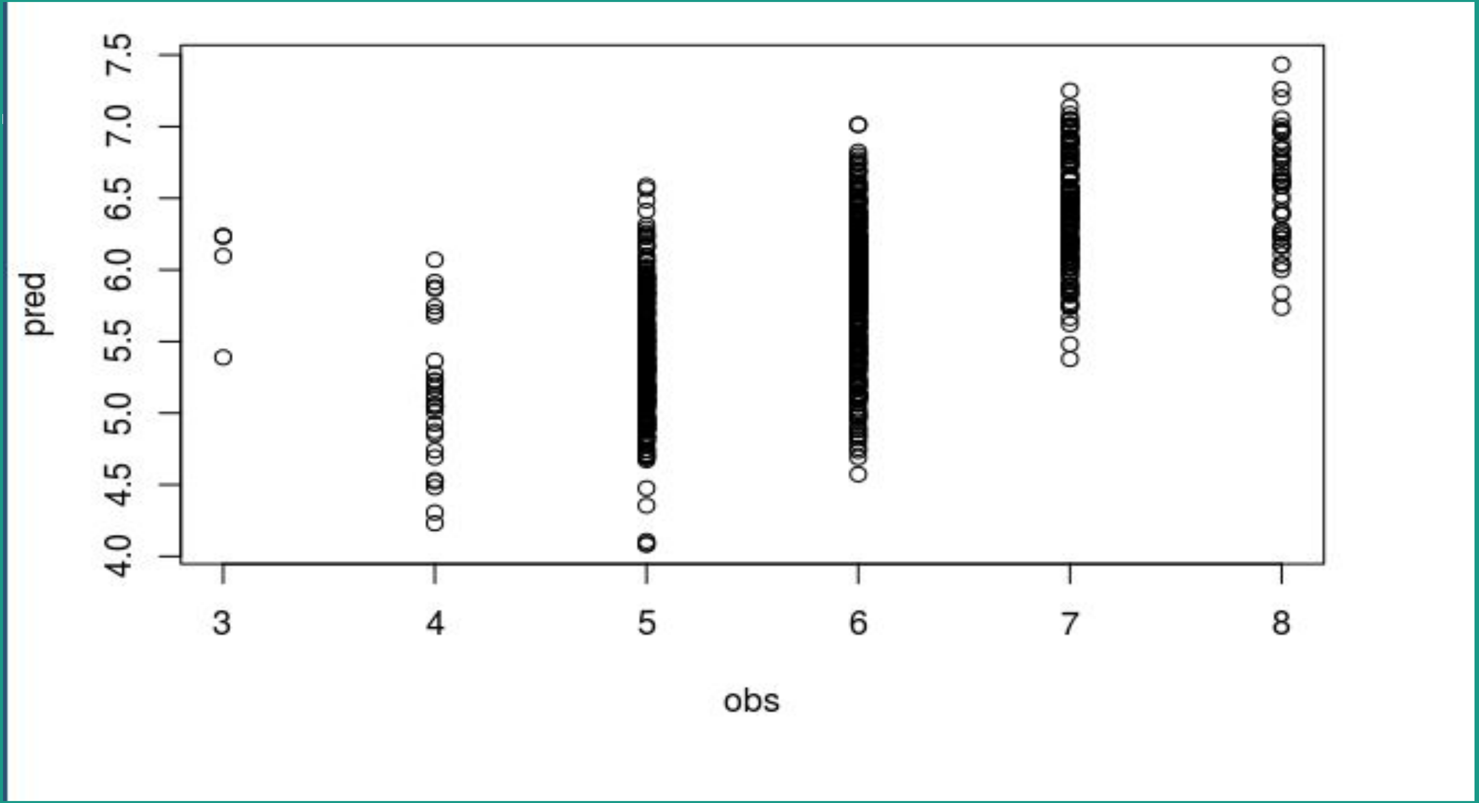
RMSE	R^2	MAE
0.6771719	0.4309241	0.5040128

Support Vector Machine



Tuning the cost parameter

Support Vector Machine



Observed vs Predicted on the test data

Summary of Models (Training Set)

Model	RMSE	RSquared	Tuning Parameters
Ordinary Linear Regression	0.7583111	0.2731315	NA
Partial Least Squares	0.7560663	0.2764547	Number of components = 8
Ridge Regression	0.7554437	0.2765135	Lambda = 0.01428571
Elastic Net	0.7515978	0.2776304	Fraction = 0.9736842 and lambda = 0.01
Lasso	0.7561563	0.2752103	Fraction = 0.7884211
K Nearest Neighbour	0.7131459	0.3534372	k = 9
Neural Network	0.6949673	0.3822180	size =9 and decay = 0.1
MARS model	0.7205307	0.3413517	degree =2 and nPrune = 18
Support Vector Machine	0.6966971	0.3816269	sigma = 0.07934471, Cost = 2

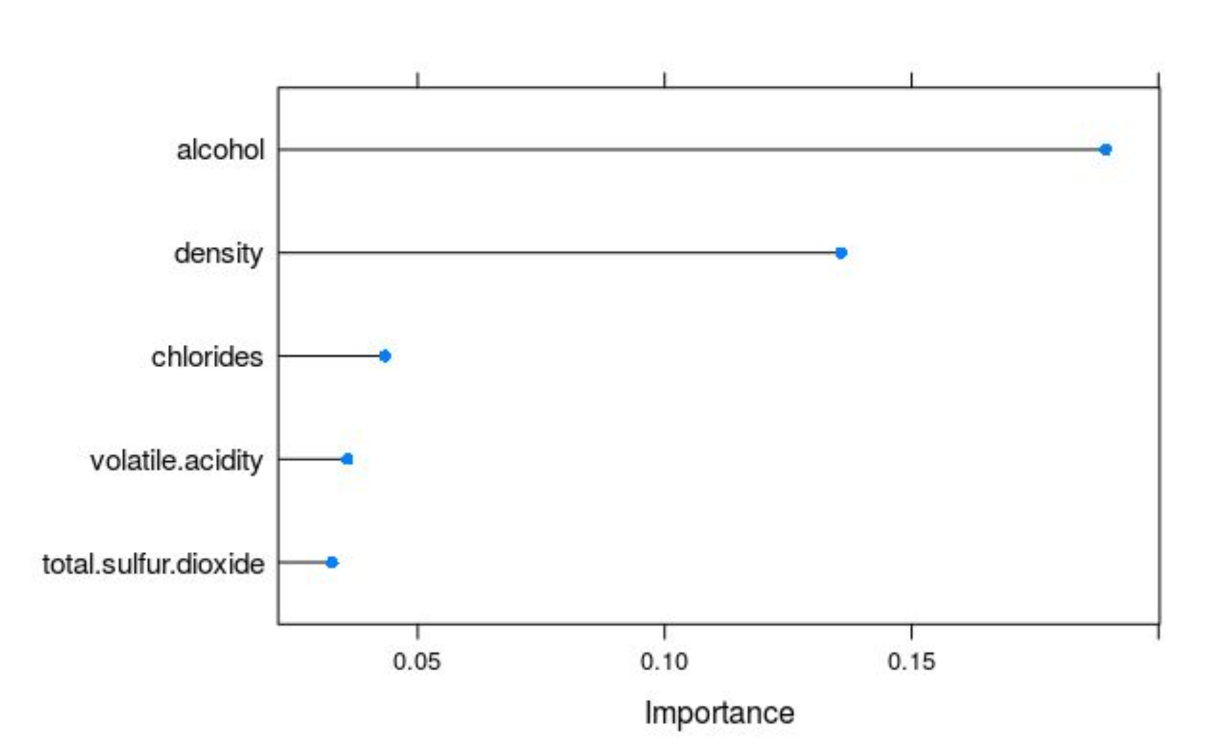
Summary of Models (Test Set)

Model	RMSE	RSquared
Ordinary Linear Regression	0.7507707	0.2684493
Partial Least Squares	0.7434357	0.2827261
Ridge Regression	0.7443214	0.2810610
Elastic Net	0.7616966	0.2749985
Lasso	0.7451390	0.2797990
K Nearest Neighbour	0.7092615	0.3741927
Neural Network	0.6790271	0.4234195
MARS model	0.7258565	0.3252741
Support Vector Machine	0.6771719	0.4309241

Result Analysis

- Neural Network and SVM the top two predictors in both the training set and testing set
- In all the cases the non linear models outperformed the linear model
- SVM has the best predictive ability among all the models
- SVM Chosen as the final model

Important Predictors



Important Variables given by SVM

Predictors	Importance
alcohol	0.1892681
density	0.1357563
chlorides	0.0434155
volatile.acidity	0.0357482
total.sulfur.dioxide	0.0327075

Conclusion and Future Work

- All the non linear models show better performance than linear models
- Non linear relationship exists between the predictors and response variable
- SVM Chosen as the final model with a highest RMSE value of 0.6771719 and RSquared value of 0.4309241 on the test set
- Classification models to be built on the data and compared with the regression model

Predicting the Quality of White Wine

Questions?

Michigan Technological University