Metro Interstate Traffic Volume

Chen Lou and Josef Pishek (Group 4)
MA 5790 Predictive Modeling

UCI Machine Learning Repository Dataset of Hourly Minneapolis-St Paul, MN traffic volume for westbound I-94. Includes weather and holiday features from 2012-2018.

Outline

Background: Automatic Traffic Recorders
Objective
Initial Data Structure
Preprocessing
Response Variables
Splitting Data
Training Resampling
Regression Models
Classification Models
Conclusion
References

Automatic Traffic Recorders (ATR)

One of several methods for collecting data on traffic volume.

Permanent installations with varying levels of technology to continuously monitor traffic volume, additional types of data depending upon their equipment and sensors.

This project’s data is from one of 70+ active devices in Minnesota-- 30+ in the seven-county metro area and 35+ in greater Minnesota.

Find more information about [Traffic Forecasting and Analysis from the Minnesota Department of Transportation (MNDoT).]

Objective

Predict the response variable "traffic_volume" from a collection of numerical and categorical predictors by:

- Preprocessing Data
- **Fitting and Evaluating Predictive Models**
  - **Regression**
  - **Classification**

Initial Data Structure

holiday	Categorical US National holidays plus regional holiday, Minnesota State Fair
temp	Numeric Average temp in kelvin
rain_1h	Numeric Amount in mm of rain that occurred in the hour
snow_1h	Numeric Amount in mm of snow that occurred in the hour
clouds_all	Numeric Percentage of cloud cover
weather_main	Categorical Short textual description of the current weather
weather_description	Categorical Longer textual description of the current weather
date_time	DateTime Hour of the data collected in local CST time
*traffic_volume	Numeric Hourly I-94 ATR 301 reported westbound traffic volume

- Categorical variables:
  - holiday (12 levels)
  - weather_main (11 levels)
  - weather_description (38 levels)
- Numerical variables:
  - temp
  - rain_1h
  - snow_1h
  - clouds_all
  - *traffic_volume
- Date and Time variables:
  - date_time
- Sample size: 9 variables with 48,204 observations
- *Response variable: traffic_volume
- Source notes: MNDoT ATR station 301, roughly midway between Minneapolis and St Paul, MN. Weather data from OpenWeatherMap.

Preprocessing - Overview

Initial Preprocessing Observations

Filtering Duplicate Observations

- There are 40,575 unique date_time values of the initial 48,204 observations

Remove “nearZeroVariance” Predictors

- holiday
- rain_1h
- Snow_1h

Remove 10 Observations for unreasonable temp values

- 40,565 observations left

Dummy Variables for Categorical Predictors

- weather_main, weather_description, year, month, day of week, hour
- With temp and clouds_all (98 Predictors)

Further Preprocessing Observations

Remove **year** and **month** Predictor

- Shows little variation with response
- Injects noise into future predictions

Remove **weather_description** Predictor

- Noisy, sparse, correlated, and summarized by weather_main

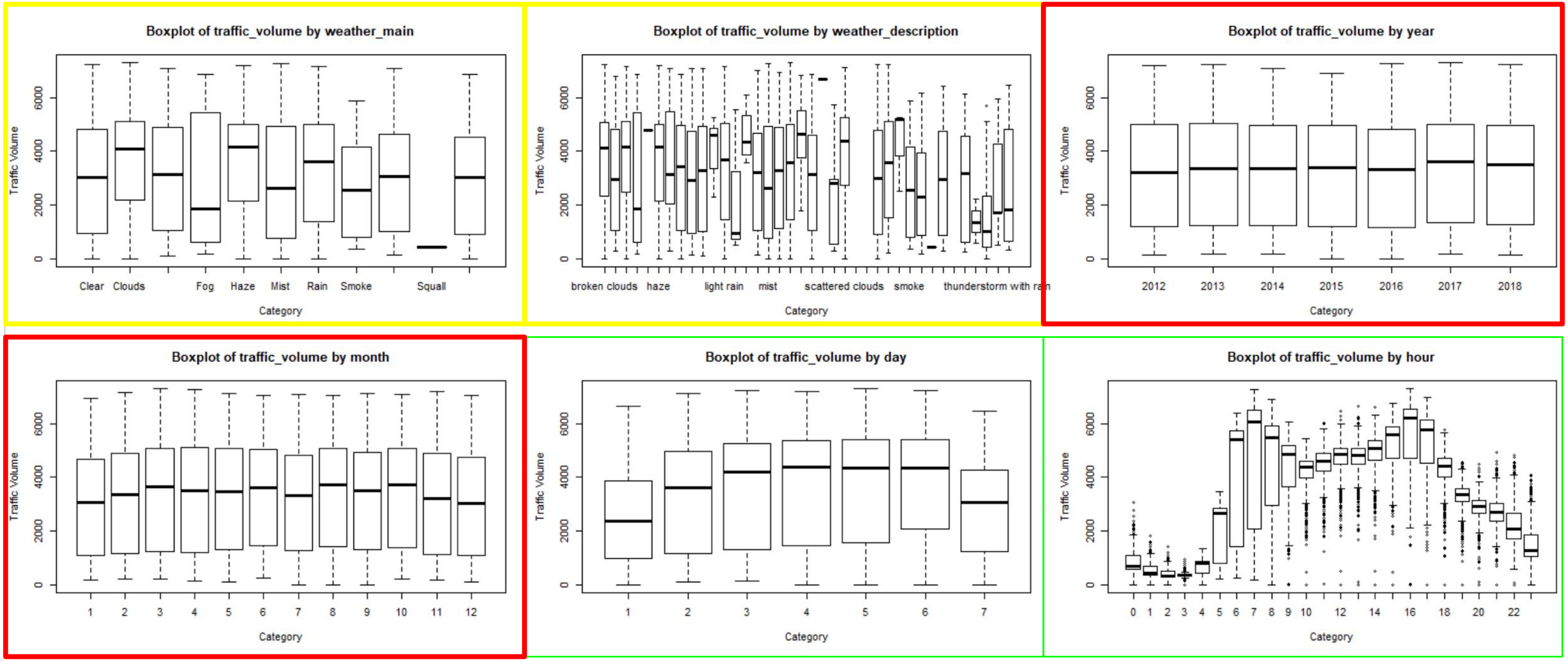
“nearZeroVariance” needs to be repeated with dummy variables (freqCut = 25/1)

- Maintains **hour**
- Removes several from **weather_main_...**

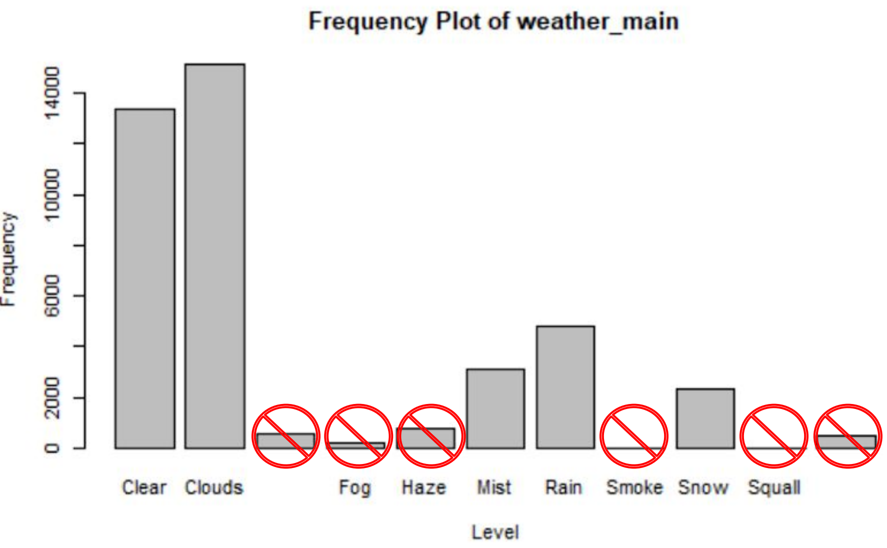
“findCorrelation” needs to be repeated with dummy variables (cutoff = .3).

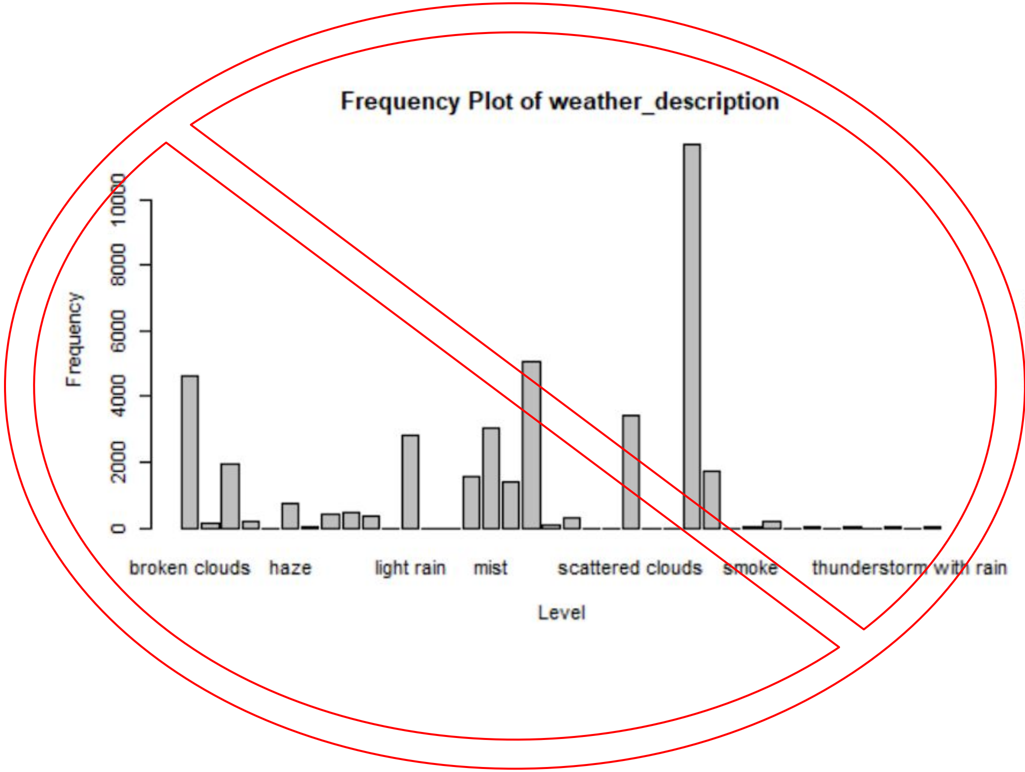
- Removes **clouds_all** and **weather_main_Clear** (keeps weather_main_Clouds) (36 Predictors)

Preprocessing - Categorical Predictors by Response

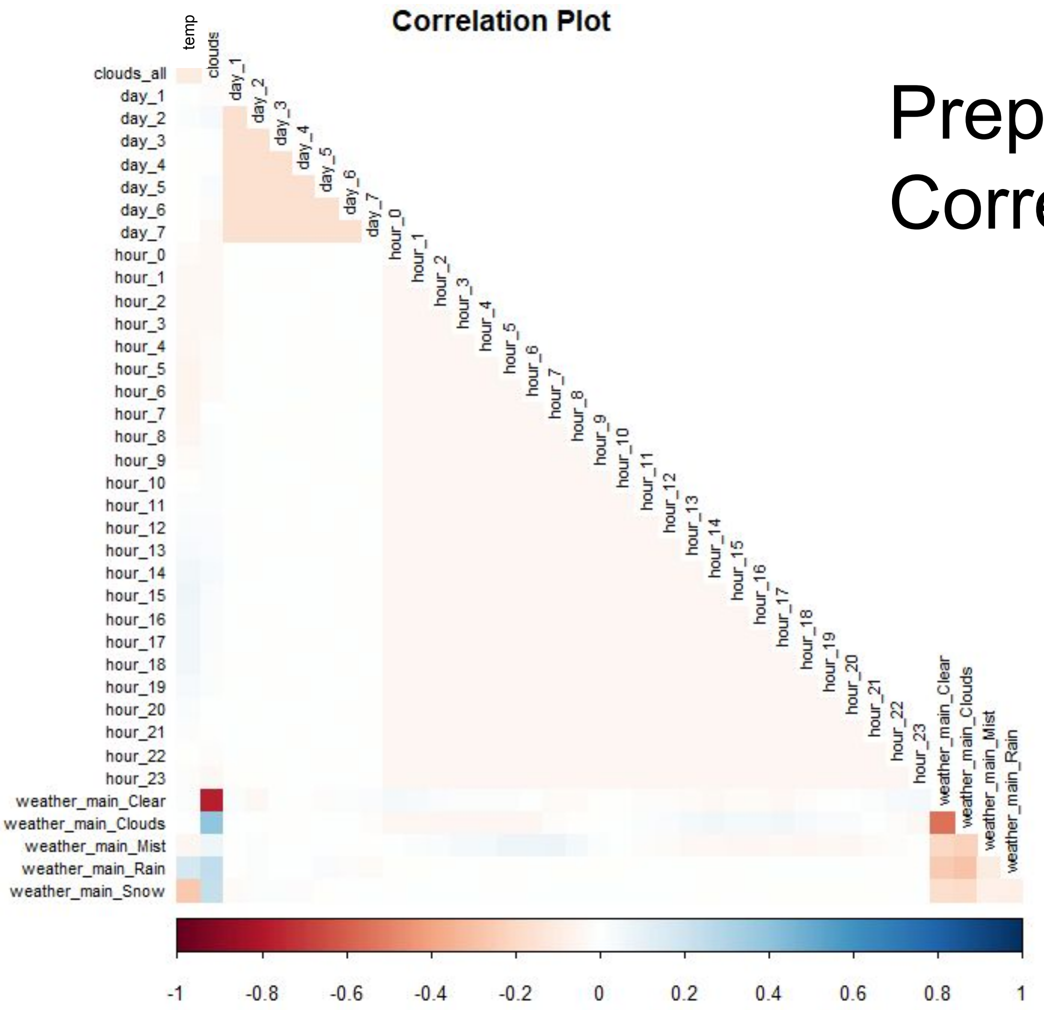


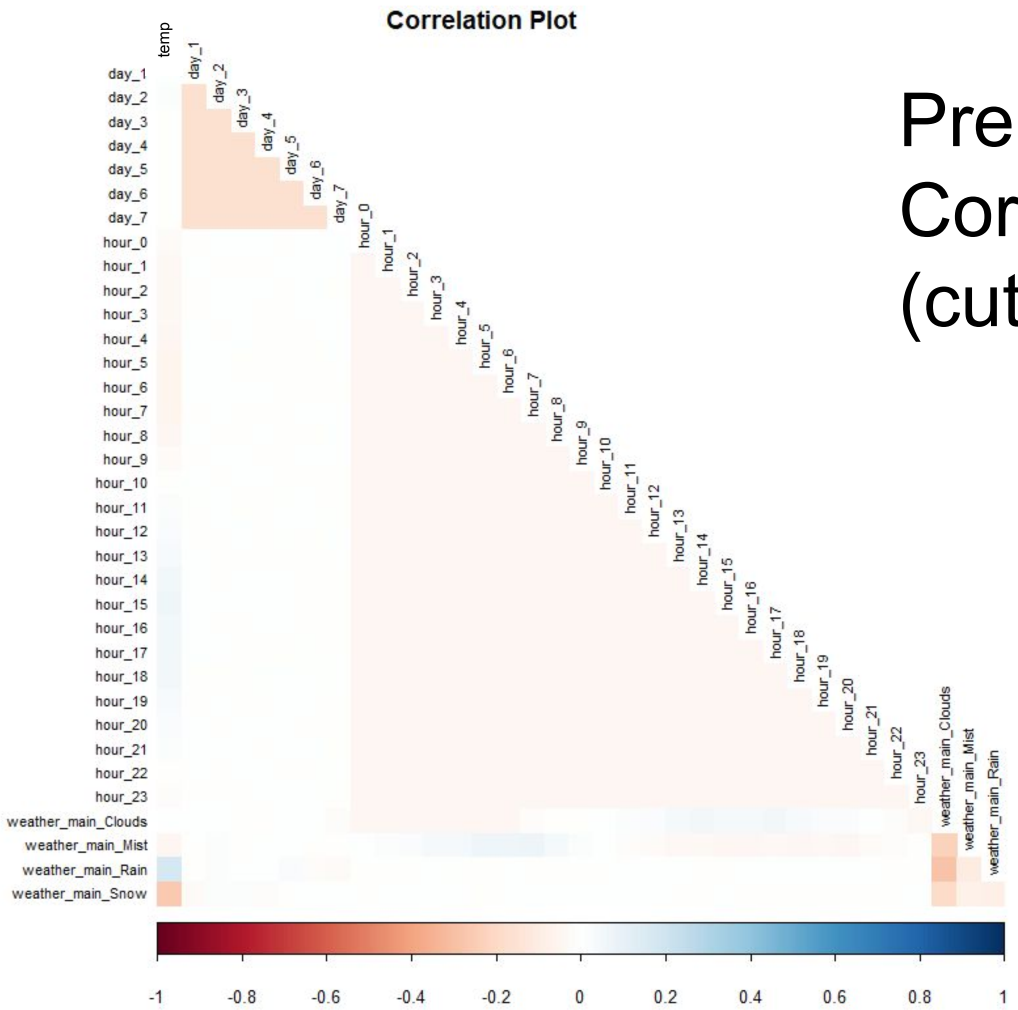
Preprocessing - Near Zero Variance Predictors





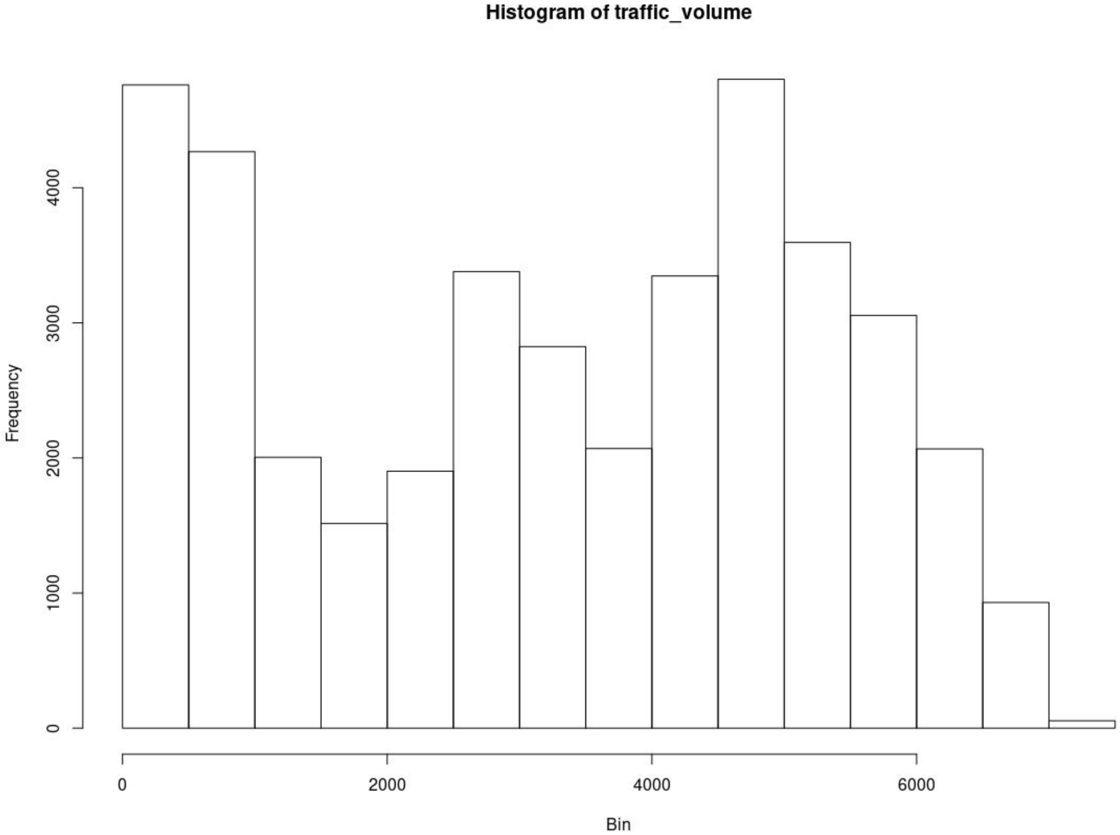
Preprocessing - Correlated Predictors





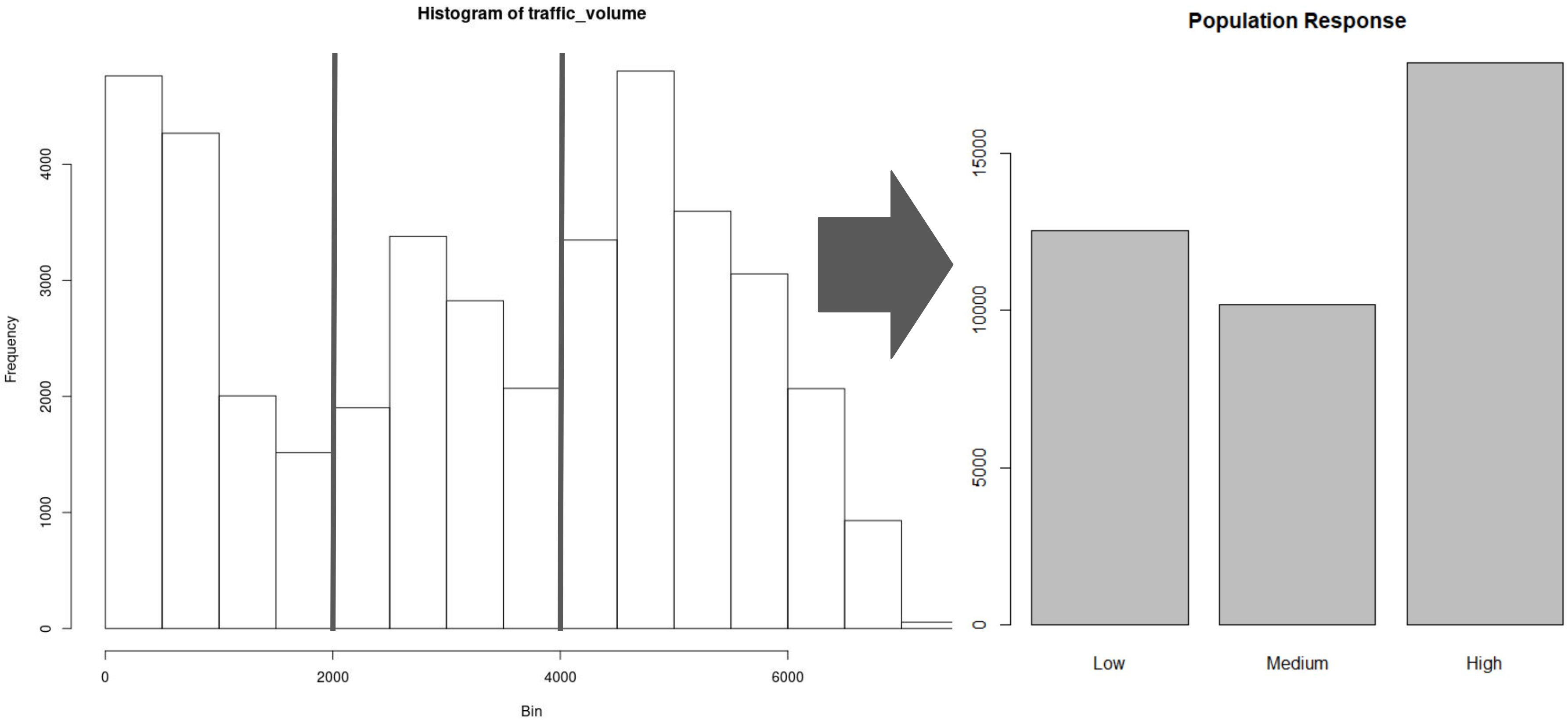
Preprocessing - Correlated Predictors (cutoff = 0.3)

Regression Response Variable

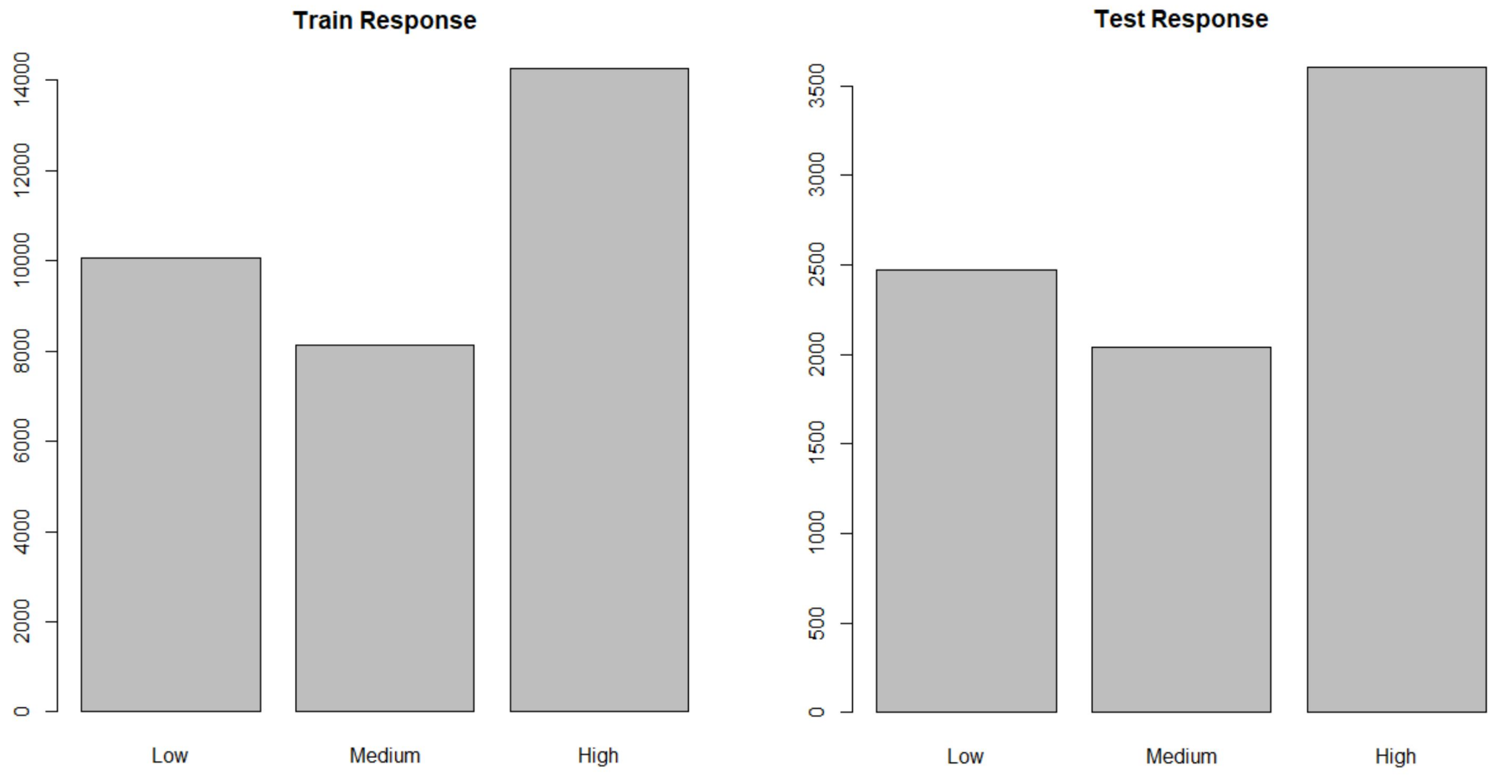


	traffic_volume
mean	3290.6505
median	3427
sd	1984.7729
0%	0
25%	1248.5
50%	3427
75%	4952
100%	7280
IQR	3703.5
skewness	-0.1074

Categorical Response Variable



Categorical Response Variable



Data Splitting

Data has been split into the following dimensions:

trainX has **32452** observations, and **36** predictors

trainY has **32452** observations

testX has **8113** observations, and **36** predictors

testY has **8113** observations

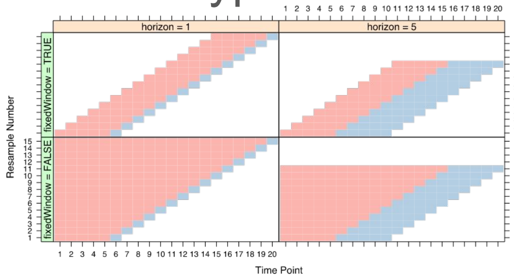
The first 80% of observations will be used for training, and the last 20% will be held-out for testing

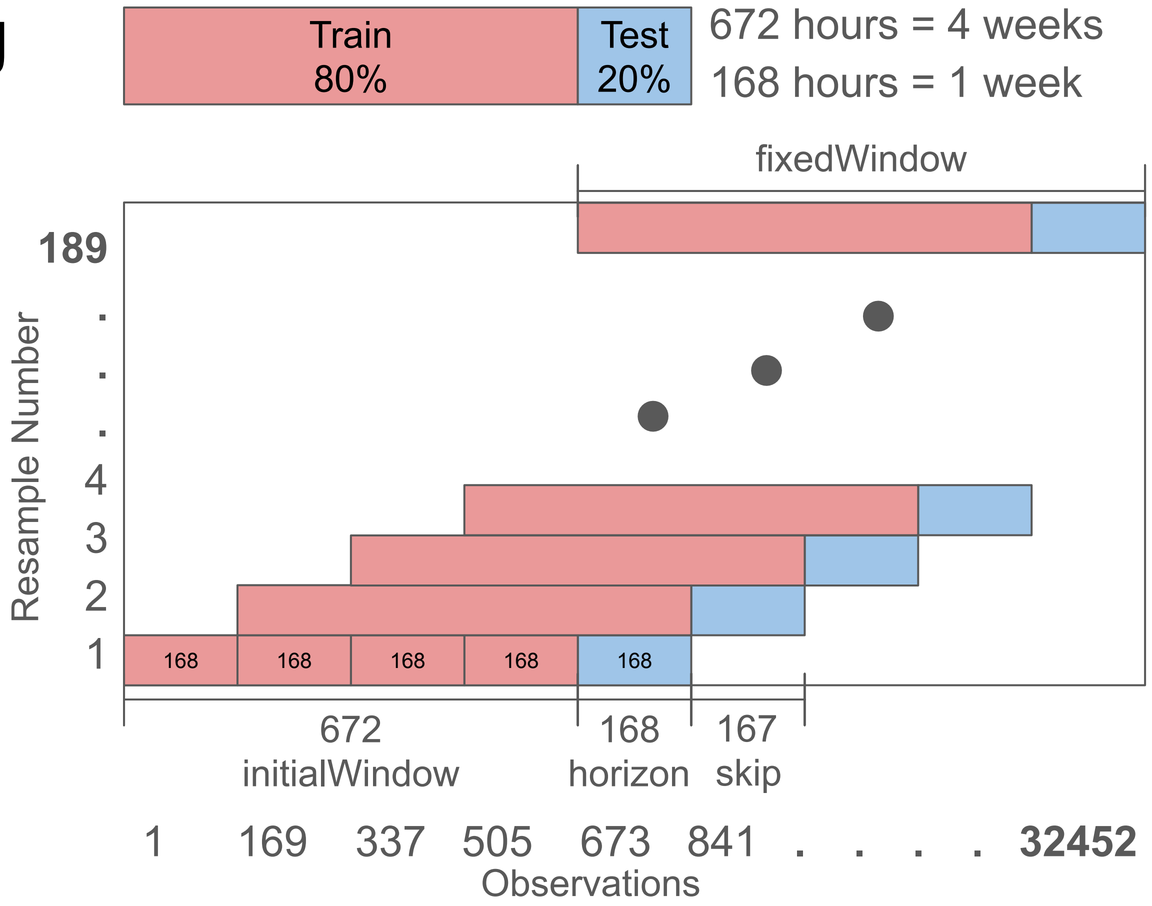
- Maintain Chronological Order
- Random splitting produces poor results for time-series data
  - More realistic predictive approach
- Essentially use traffic_volume from 2012-2017 to predict 2018

Training Resampling

Rolling Forecasting Origin - “Timeslice” Method:

- Parameters
  - fixedWindow = **TRUE**,
  - horizon = **168**,
  - initialWindow = **672**,
  - skip = **167**
- These were chosen for computational efficiency, several types available:



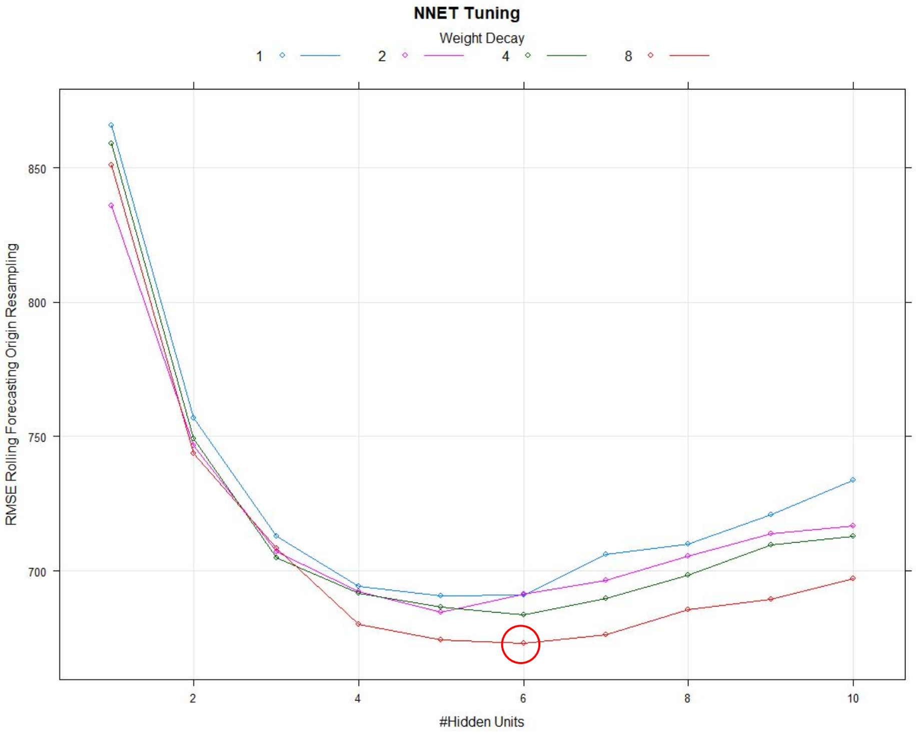


Regression Performance Summary

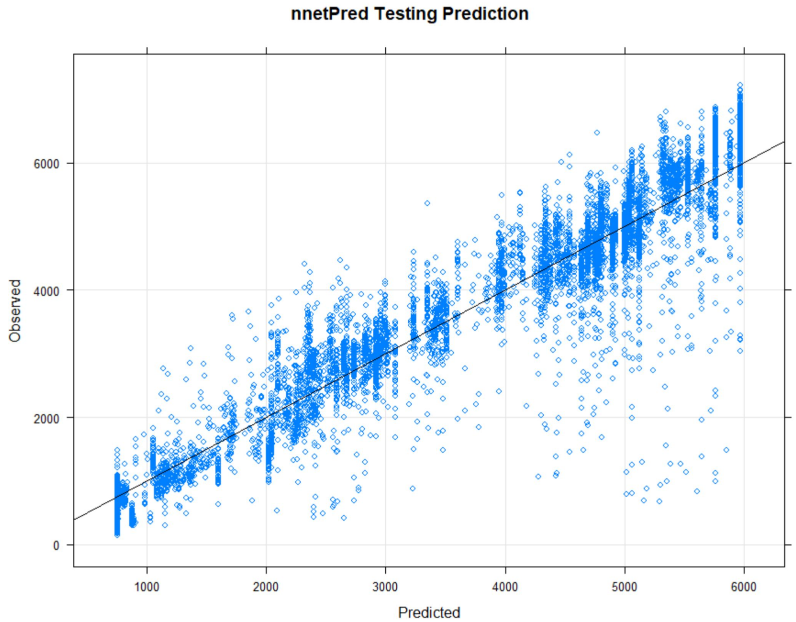
Model	Train RMSE	Train R^2	Train Time (sec)	Test RMSE	Test R^2
PLS	814.33	0.84	3	792.56	0.84
ENET	834.94	0.83	68	794.17	0.84
LARS	813.24	0.84	5	792.56	0.84
NNET	673.14	0.89	13797	569.76	0.92
MARS	617.58	0.90	843	548.61	0.92
SVM	820.23	0.84	2412	**476.93**	**0.94**
KNN	**584.25**	**0.91**	**57**	544.60	0.92

Neural Network Regression

Model	Train RMSE	Train R^2	Train Time (sec)	Test RMSE	Test R^2
NNET	673.14	0.89	13797	569.76	0.92

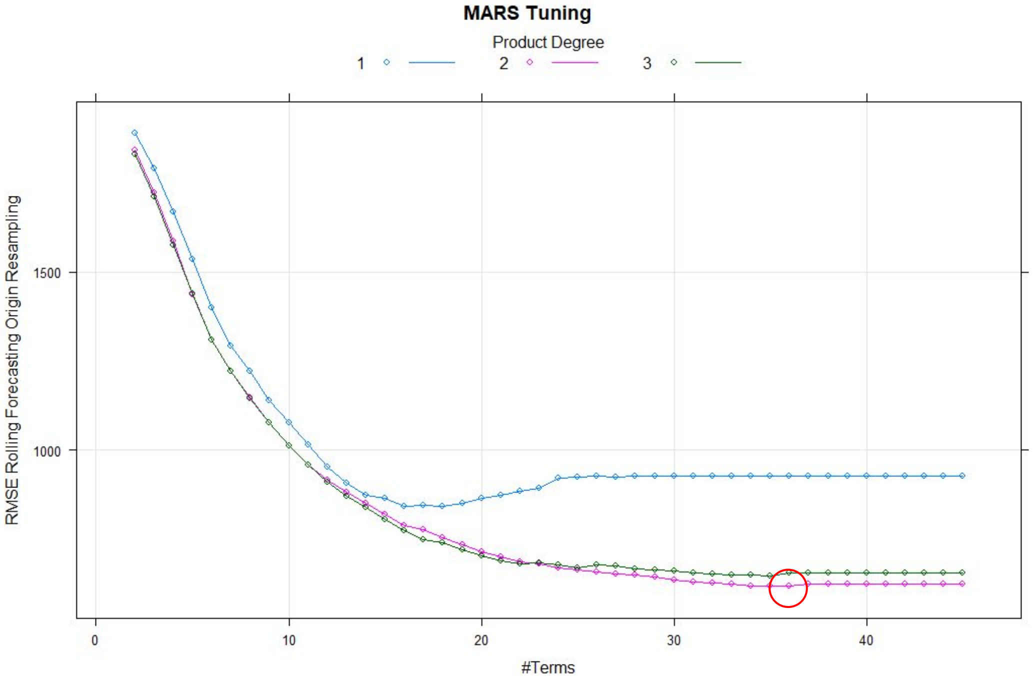



```
Tuning parameter 'bag' was held constant at a value of FALSE
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were size = 6, decay = 8 and
 bag = FALSE.
```

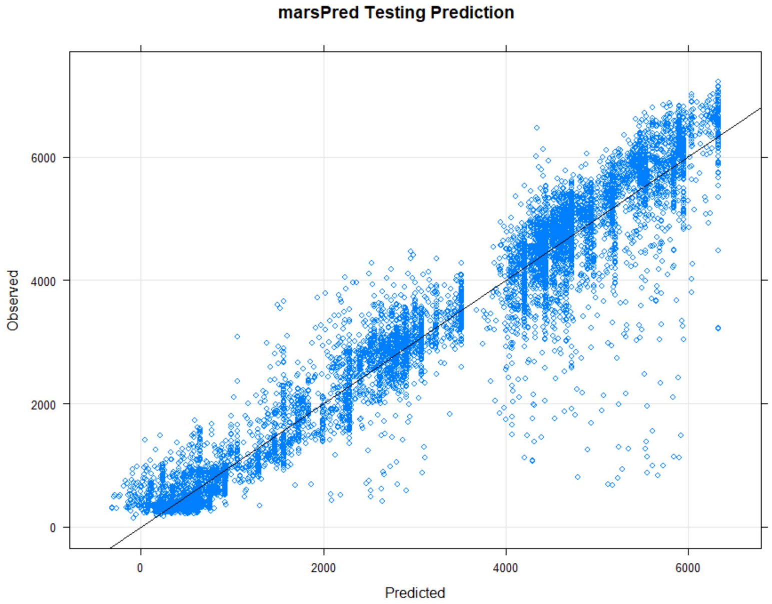


Multivariate Adaptive Regression Splines

Model	Train RMSE	Train R^2	Train Time (sec)	Test RMSE	Test R^2
MARS	617.58	0.90	843	548.61	0.92

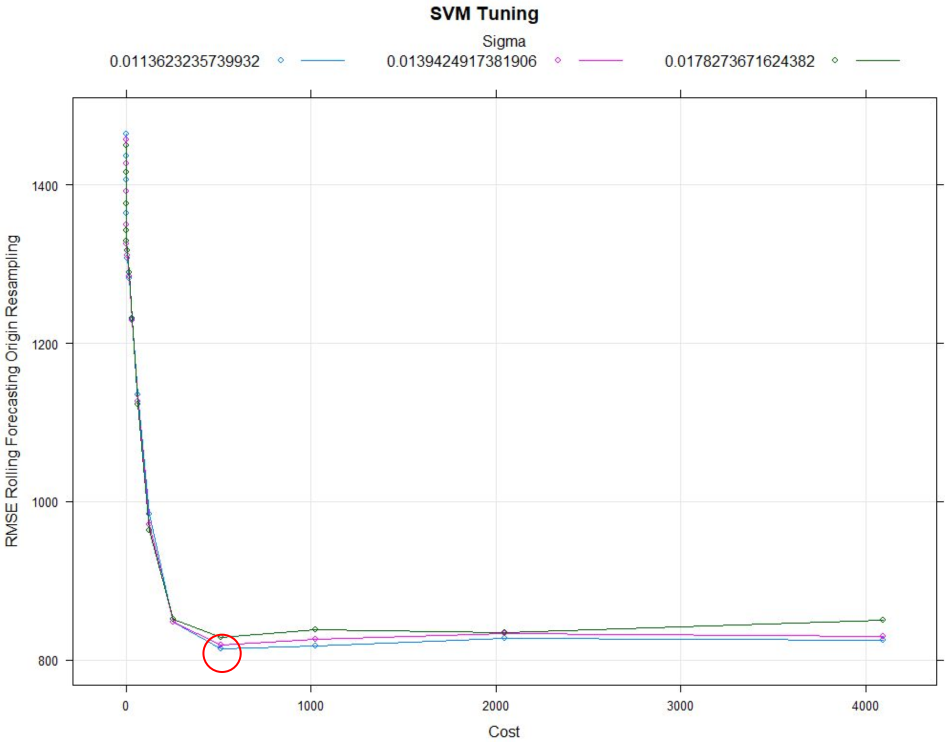



```
RMSE was used to select the optimal model using
 the smallest value.
The final values used for the model were nprune = 36
 and degree = 2.
```

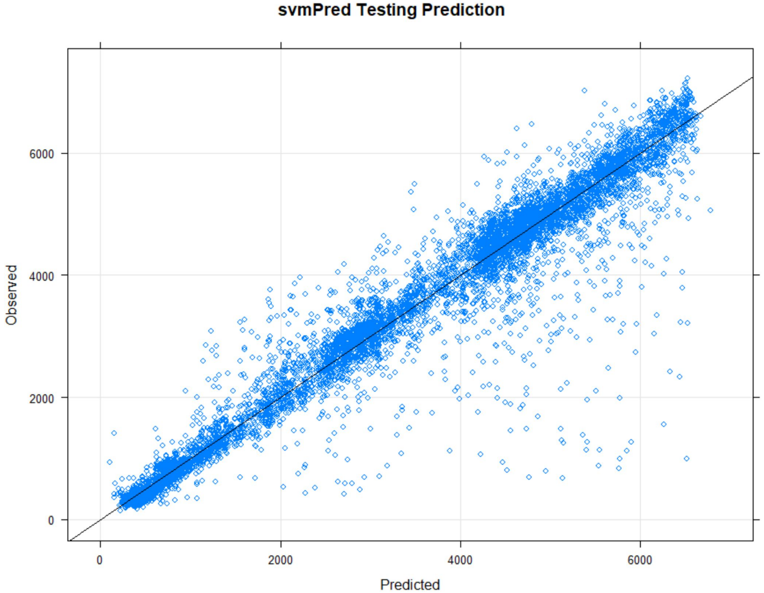


Support Vector Machines Regression

Model	Train RMSE	Train R^2	Train Time (sec)	Test RMSE	Test R^2
SVM	814.06	0.84	2432	**475.43**	**0.94**

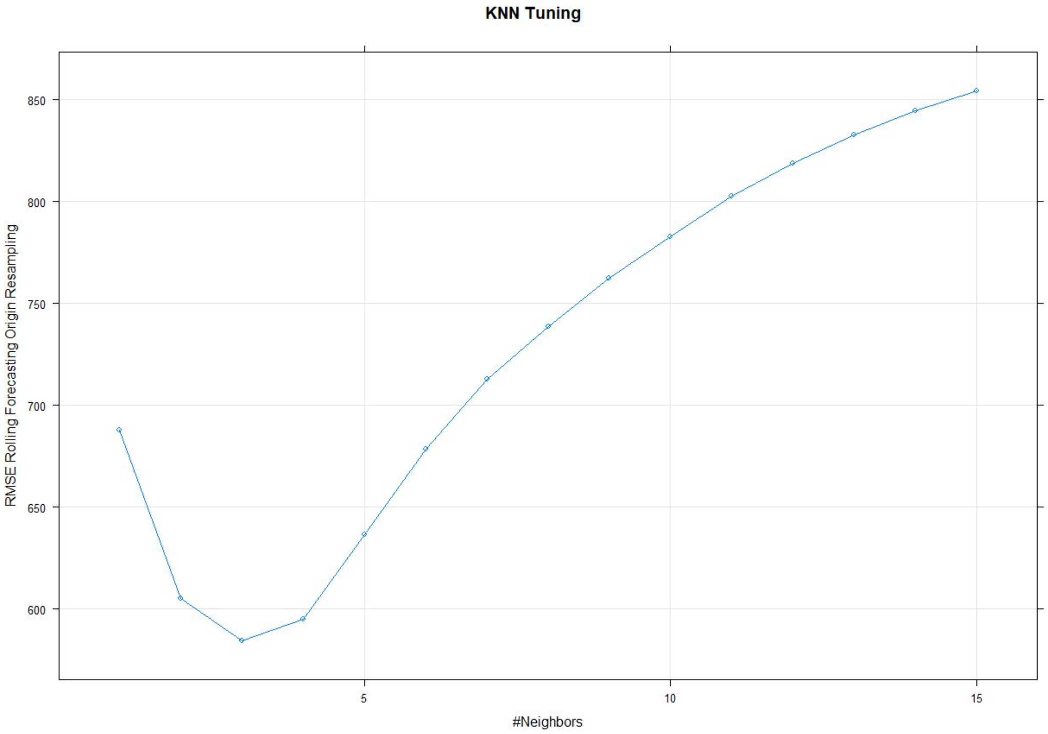



```
RMSE was used to select the optimal model using the
 smallest value.
The final values used for the model were sigma =
 0.01136232 and C = 512.
```

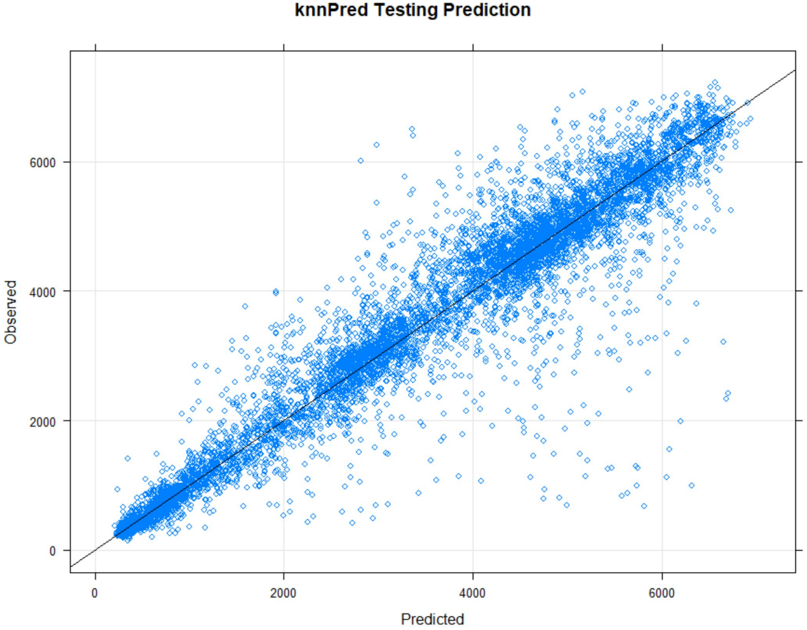


K Nearest Neighbors Regression

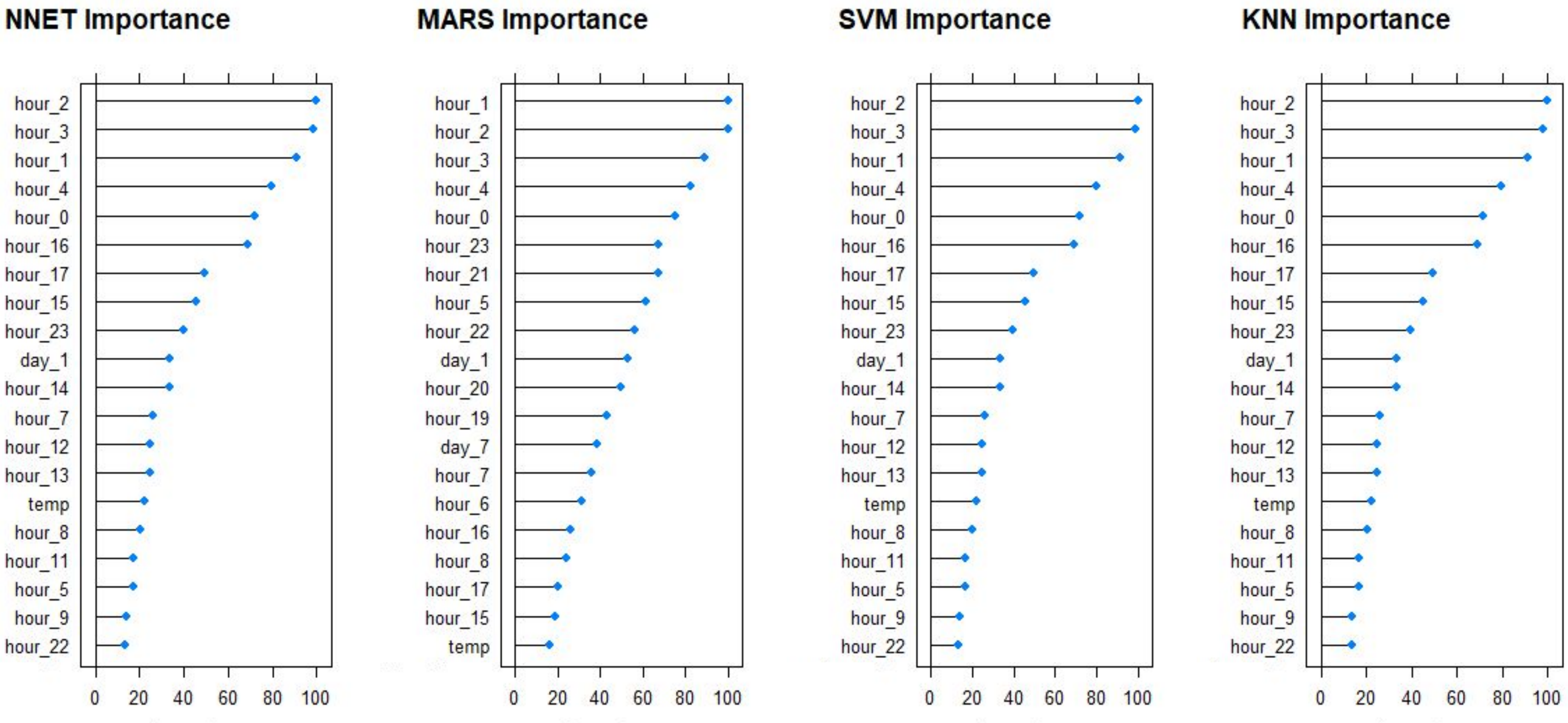
Model	Train RMSE	Train R^2	Train Time (sec)	Test RMSE	Test R^2
KNN	**584.25**	**0.91**	57	544.60	0.92

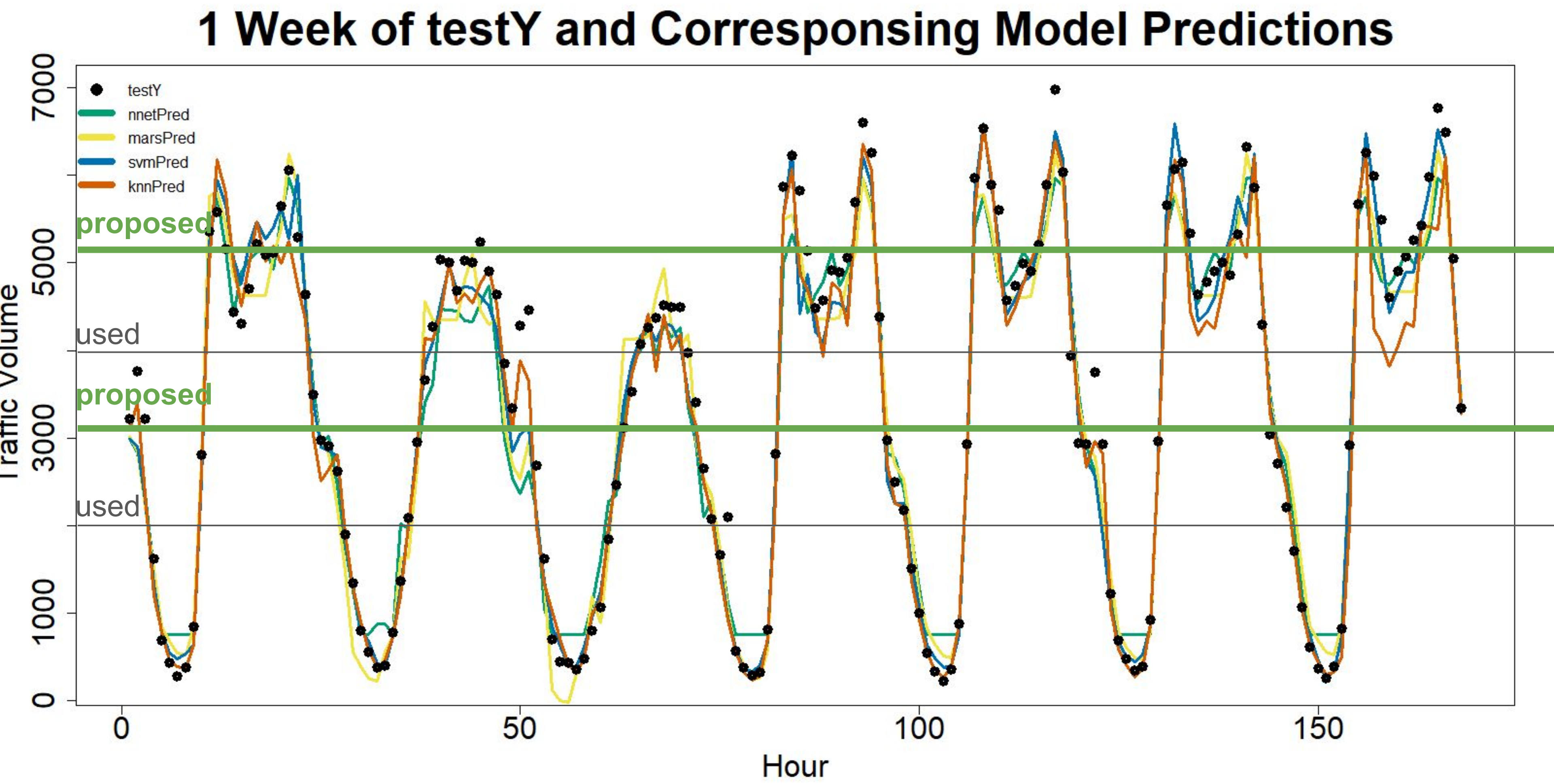



```
RMSE was used to select the optimal model using
 the smallest value.
The final value used for the model was k = 3.
```



Regression Variable Importance



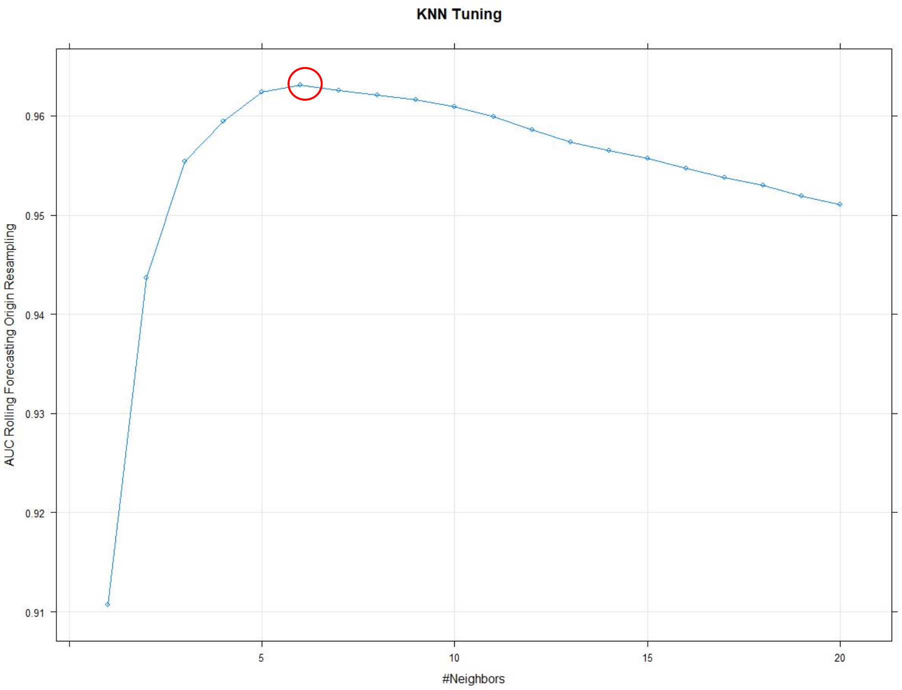


1 Week of testY and Corresponsing Model Predictions
testY
nnetPred
marsPred
svmPred
knnPred
proposed
used
proposed
used
Traffic Volume
0
1000
3000
5000
7000
0
50
100
150
Hour

Regression Performance Summary

Model	Train RMSE	Train R^2	Train Time (sec)	Test RMSE	Test R^2
PLS	814.33	0.84	3	792.56	0.84
ENET	834.94	0.83	68	794.17	0.84
LARS	813.24	0.84	5	792.56	0.84
NNET	673.14	0.89	13797	569.76	0.92
MARS	617.58	0.90	843	548.61	0.92
SVM	820.23	0.84	2412	**476.93**	**0.94**
KNN	**584.25**	**0.91**	**57**	544.60	0.92

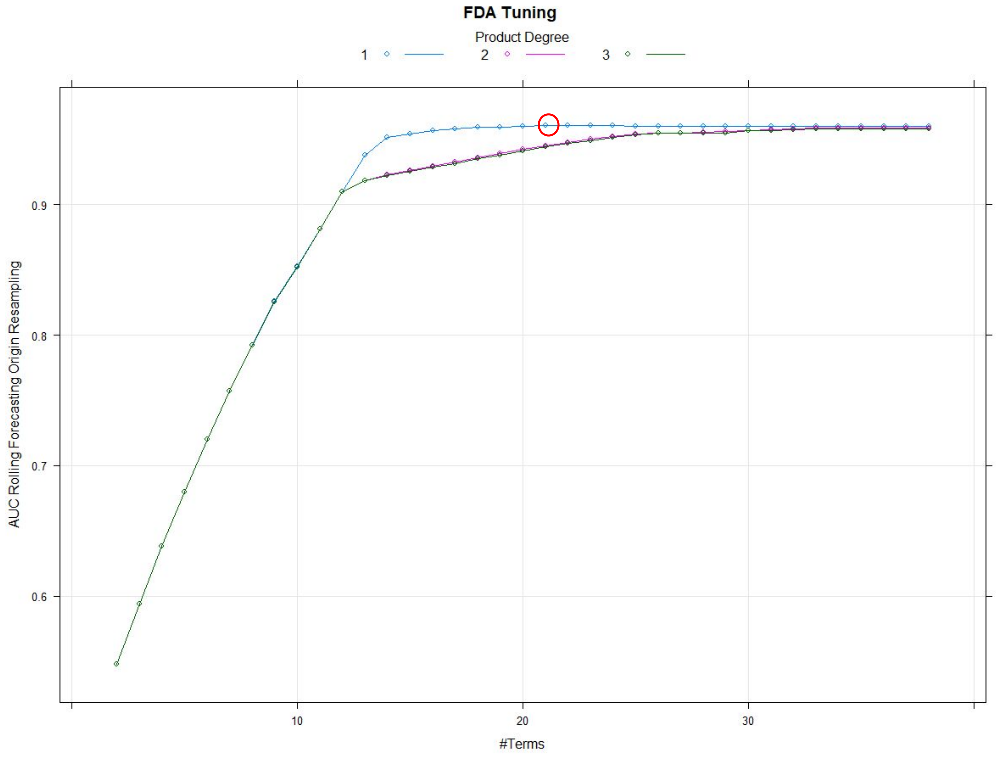
KNN Classification




```
Confusion Matrix and Statistics

          Reference
Prediction  Low Medium High
    Low    2337    113    0
    Medium  104   1677  129
    High     33    248 3472

Overall Statistics

               Accuracy : 0.9227
                 95% CI : (0.9167, 0.9284)
    No Information Rate : 0.4439
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8799

 Mcnemar's Test P-Value : 2.691e-15

Statistics by Class:
```

```
AUC was used to select the optimal model using the largest value.
The final value used for the model was k = 6.
```

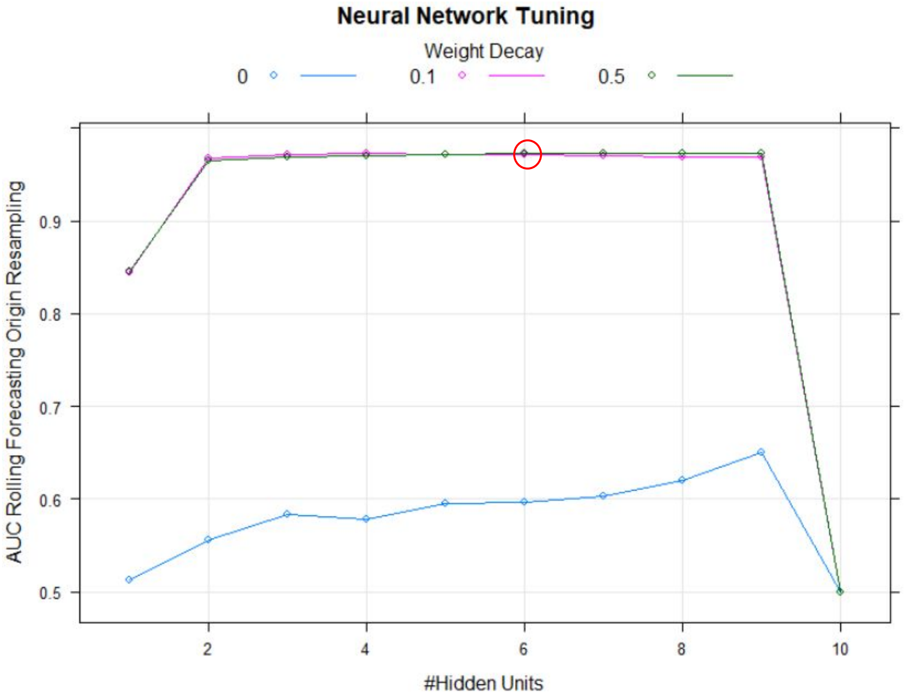
FDA Classification




```
Confusion Matrix and Statistics

          Reference
Prediction  Low Medium High
    Low    1995     87    0
    Medium  276   1425   20
    High    203    526 3581

Overall Statistics

               Accuracy : 0.8629
                 95% CI : (0.8553, 0.8703)
    No Information Rate : 0.4439
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7834

 Mcnemar's Test P-Value : < 2.2e-16
```

```
AUC was used to select the optimal model using the largest value.
The final values used for the model were degree = 1 and nprune = 21.
```

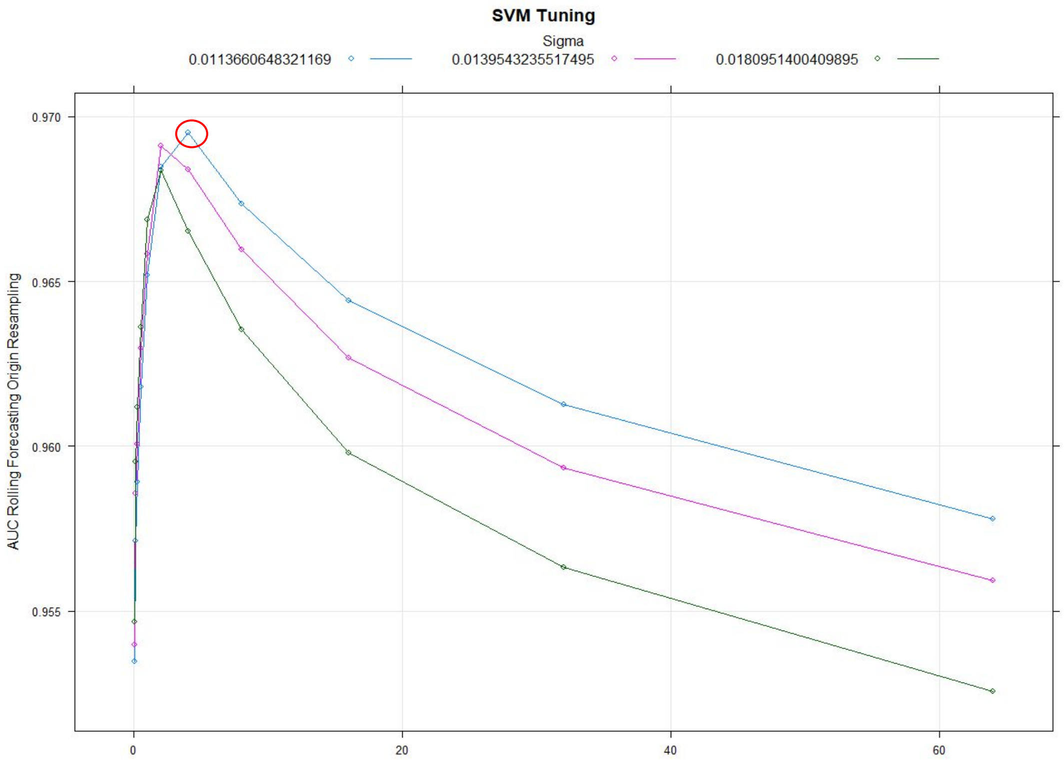
Neural Network Classification




```
Confusion Matrix and Statistics

          Reference
Prediction  Low Medium High
    Low    2357    134    0
    Medium   82   1630   85
    High     35    274 3516

Overall Statistics

               Accuracy : 0.9248
                 95% CI : (0.9189, 0.9305)
    No Information Rate : 0.4439
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8828

 Mcnemar's Test P-Value : < 2.2e-16
```

```
Tuning parameter 'bag' was held constant at a value of FALSE
AUC was used to select the optimal model using the largest value.
The final values used for the model were size = 6, decay = 0.5 and bag = FALSE.
```

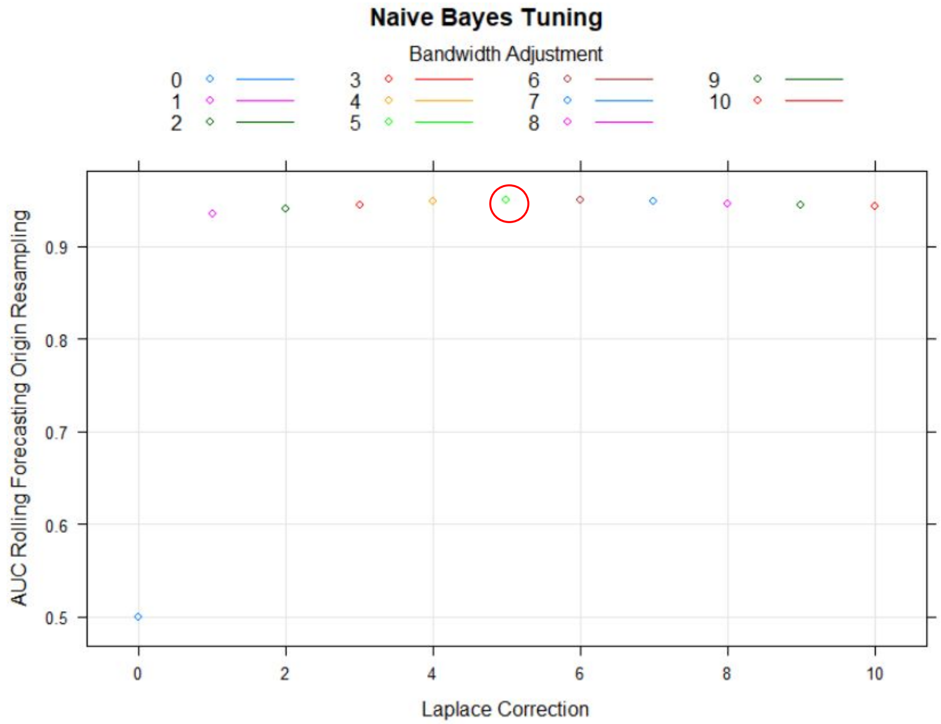
SVM Classification




```
Confusion Matrix and Statistics

          Reference
Prediction  Low Medium High
    Low    2355    129    0
    Medium   84   1627   68
    High     35    282 3533

Overall Statistics

               Accuracy : 0.9263
                 95% CI : (0.9204, 0.9319)
    No Information Rate : 0.4439
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.885

 Mcnemar's Test P-Value : < 2.2e-16
```

```
AUC was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.01136606 and C = 4.
```

Naive Bayes Classification



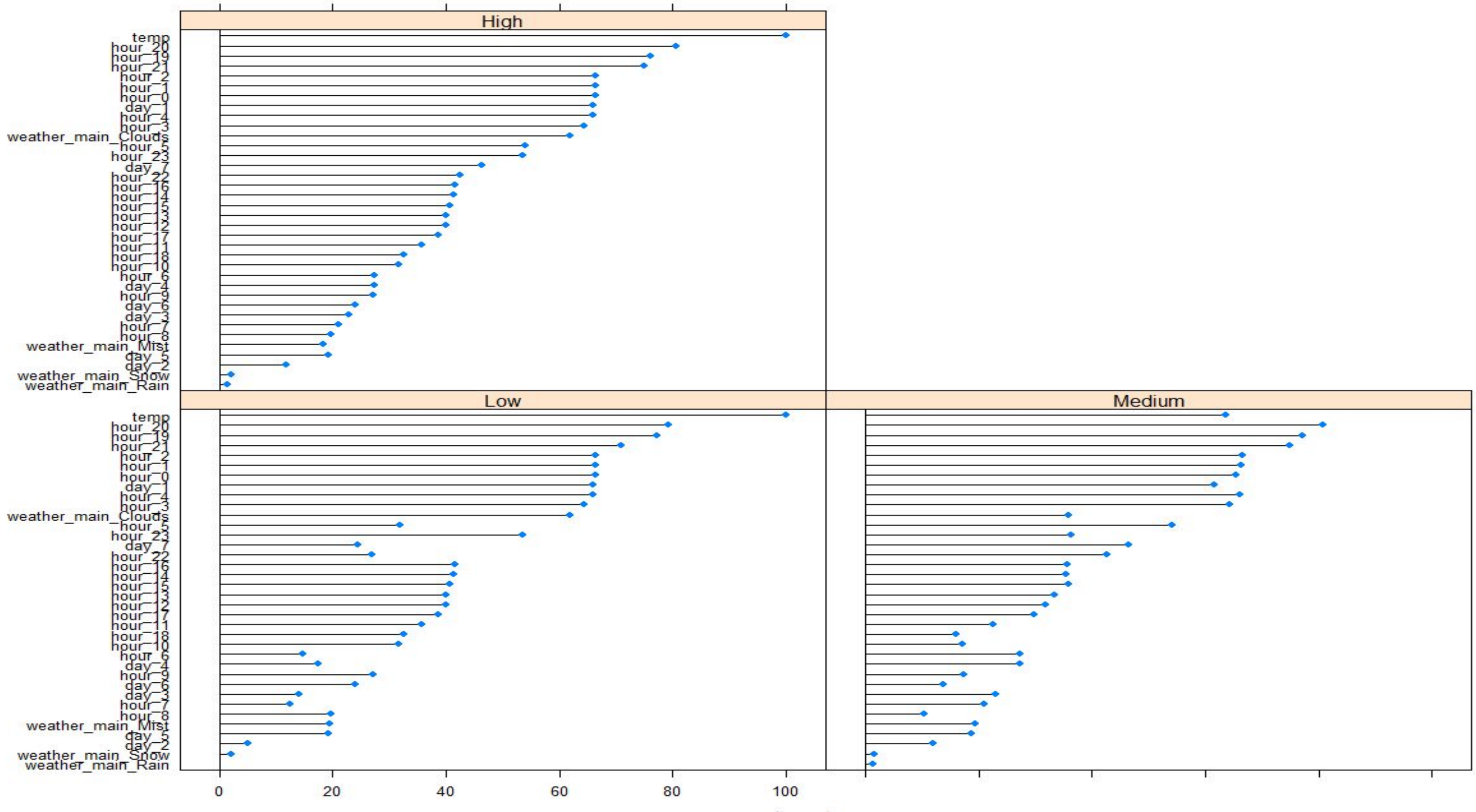

```
Confusion Matrix and Statistics

          Reference
Prediction  Low Medium High
    Low    2463   1847 1209
    Medium    0      0    0
    High     11    191 2392

Overall Statistics

               Accuracy : 0.5984
                 95% CI : (0.5877, 0.6091)
    No Information Rate : 0.4439
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3828

 Mcnemar's Test P-Value : < 2.2e-16
```

```
Tuning parameter 'usekernel' was held constant at a value of TRUE
AUC was used to select the optimal model using the largest value.
The final values used for the model were fL = 5, usekernel = TRUE and adjust = 5.
```

Classification Summary

Model	Train Accuracy	Train Kappa	Train Time* (sec)	Test Accuracy	Test Kappa
KNN	0.8982	0.8430	74.31	0.9227	0.8799
FDA	0.8601	0.7817	1787.49	0.8629	0.7834
Neural Network	0.9116	0.8641	3117.12	0.9248	0.8828
SVM	0.9123	0.8654	517.05	0.9263	0.8850
Naive Bayes	0.6720	0.4882	658.08	0.5984	0.3828

Classification Variable Importance



Conclusion

Best Models

- Regression: Non-linear Methods
  - KNN
  - SVM
- Classiffication: Non-linear Methods
  - NNET
  - SVM

Before over-fitting this dataset, explore data for another ATR location to see bias / variability

Questions?

References

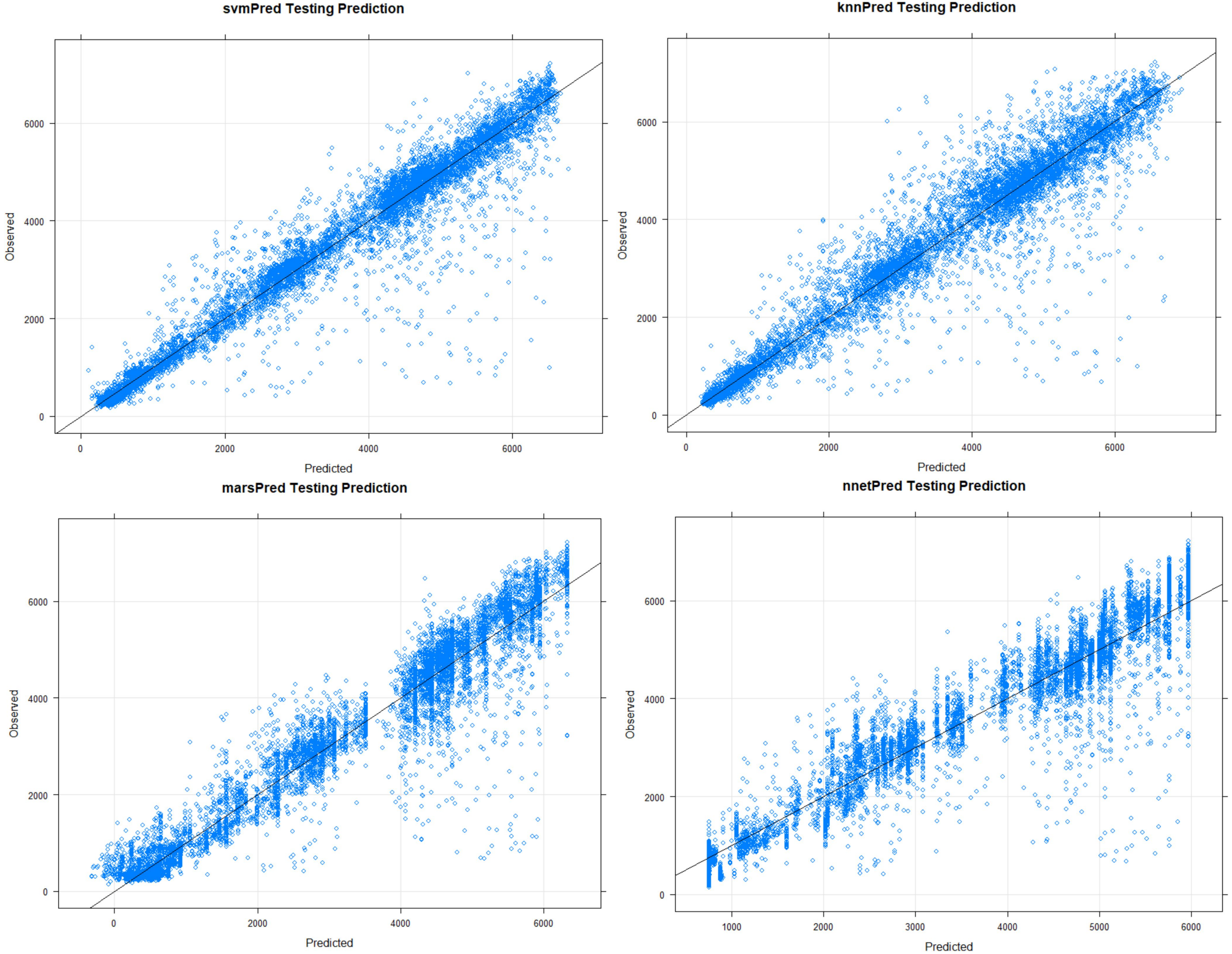
Hogue, John. (2019). “Metro Interstate Traffic Volume Data Set”. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume

Hyndman, R.J. & Athanasopoulos, G (2018). *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. https://otexts.com/fpp2/accuracy.html

Kuhn, M (2019). “Data Splitting for Time Series”. *The caret package*. https://topepo.github.io/caret/data-splitting.html#data-splitting-for-time-series

Kuhn, M & Johnson, K (2013). *Applied Predictive Modeling*. Springer Science + Business Media.

MNDoT (2019). “Collection Methods”. Traffic Forecasting and Analysis. St. Paul, MN. http://www.dot.state.mn.us/traffic/data/coll-methods.html


svmPred Testing Prediction
Observed
Predicted
0
2000
4000
6000
knnPred Testing Prediction
Observed
Predicted
0
2000
4000
6000
marsPred Testing Prediction
Observed
Predicted
0
2000
4000
6000
nnetPred Testing Prediction
Observed
Predicted
0
2000
4000
6000
1000
2000
3000
4000
5000
6000