Contents

1 Introduction

In this report, we implemented some regression models to fit the data within two datasets, including the linear regression, random forest regression, neural network regression and polynomial regression. We made some comparisons between the prediction results of different models based on root mean squared error(RMSE) and the coefficient of determination(R^2). Based on the statistics provided by models, we analyzed the significance of different features. In order to avoid overfitting, which often happens in the regression fitting problem, we implemented both cross-validation and regularization techniques. Moreover, since some of the features have no numerical meaning, we formatted them into binary features, which increase the number of features.

The report is organized as follows : In section 2, we preprocessed the data and plotted the data to find some pattern. In section 3, The dataset was from a simulated traffic data on the backup system in a networkwe. We implemented both the linear regression model and random forest regression model to predict the copy size of the backup file and made some analysis and comparisons of the predictions. In section 4, we implemented the neural network model and polynomial regression model to fit the data and analyzed the impact of different parameters to the results. In section 5...

2 Data Preprocessing

2.1 Feature Binarization

After checking the data, there are seven features in the dataset named network_backup_ dataset, including week, day of the week, backup start time, work flow ID, file name, size of backup and backup time. The target value is the size of backup, others are features. Therefore, first of all we need to transform non-numerical values into numerical values. However, consider "the day of the week" feature, "Tuesday" is not larger than "Monday", but "2" is larger than "1". So we need to make such kind of features into binary format. For instance, for "day of the week", we need to expand it into seven features, where it is 1 on the specified day and all the others 0. Finally, we have 64 features in total.

2.2 Pattern Detection

In order the find the pattern, we plot the figure of the copy size over time. The independent variable is the "Start Time - Hour of Day", the dependent variable is the copy size. There are six sampling times in a day and 20 days as a period. So the x axis is from 1 to 120. The figure is shown in figure 1. From 1 we can see there is clearly a cycle between copy size and time. So we plot the copy size over days to find the cycle period. The result is shown in figure 2. Hence we can predict that the cycle period is approximately a week.

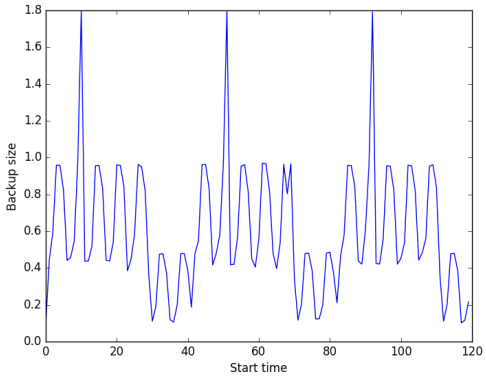

Figure 1: The copy size of backup over start time

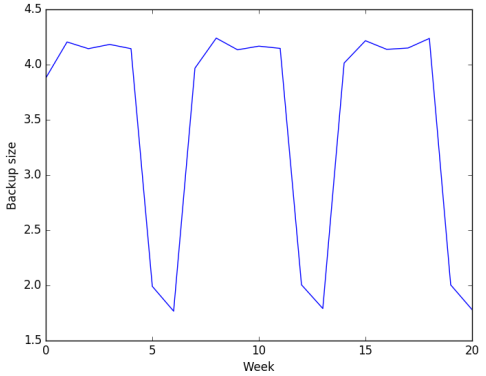

Figure 2: The copy size of backup over week

3 Linear and Random Forest Regression

In the Linear Regression Model, we choose the copy size as the target variable and the other attributes as the features, the fit model is as follow

$$\bar{Y} = X\alpha$$

where $\bar{Y}$ represents the target, X represents the features vector and α is the linear coefficient. We used least square as the penalty function, which is

$$min||Y - \bar{Y}||_2$$

In order to avoid the overfitting problem, we implemented 10-fold cross-validation technique. First do a random shuffling over the data, then split them into 10 folds. Each time choose 9 of 10 as the train set and the other 1 fold as the test set. The Root Square Mean Error (RMSE) is shown in figure 3. The average RMSE is 0.071. To evaluate the fitting accuracy of our model, Fitted values

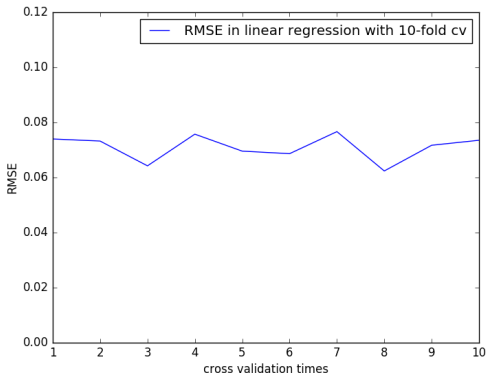

Figure 3: The copy size of backup over week

VS actual values and residuals vs fitted values are plotted. From figure 4, we can see that most predicted values are close to the actual values. From figure 3, we can see that the residuals are distributed randomly around zero axis, which means the model is proper.

As for the significance of different variables, we choose p-value as the evaluation criterion. The p-value of the six features is shown in figure 6.

Therefore the most important two features are "Backup start time - Hour of the day" and "Backup Time" because their p-values are close to 0. For 64 features, the p-value matrix is shown below.

For random forest regression, we initialize the model with parameters : number

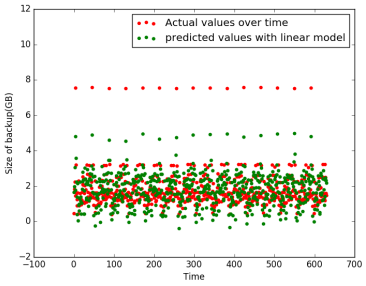

Figure 4: Actual VS predicted

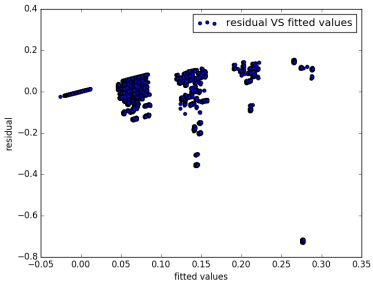

Figure 5: Residual VS fitted values

```
[  7.78443589e-01   8.71062213e-04   2.46112227e-08   1.32748780e-01
   1.57542554e-01   0.00000000e+00]
```

Figure 6: p-value of 6 features

```
7.36199210e-001   9.39383338e-001   4.66439149e-001   9.08462632e-001
5.45241529e-001   4.35994229e-001   3.09252234e-001   6.38013695e-001
5.16578878e-001   2.80070834e-001   4.85768453e-001   8.55993609e-001
6.36535181e-001   7.63174401e-001   2.98120865e-001   1.18985257e-004
9.31487572e-004   1.66246279e-001   1.95819809e-001   4.44216789e-001
5.17062882e-001   1.99647310e-024   6.99156954e-009   1.54736893e-004
1.23481345e-002   3.53017379e-014   5.81362330e-002   1.24279127e-001
1.59934030e-033   8.85184242e-006   7.79480962e-034   2.34530310e-060
2.97238246e-100   2.68855252e-006   1.23242625e-005   3.15955319e-004
1.10012870e-005   5.78513377e-005   2.40988217e-007   2.36501171e-003
3.07991642e-001   4.75428019e-002   6.68776922e-002   5.09616461e-002
8.84143253e-001   6.74472700e-005   9.82830391e-006   6.18754039e-006
7.10107131e-003   9.41370915e-007   2.92999661e-010   5.36294108e-008
2.37168301e-008   1.50927326e-011   4.51686551e-009   2.17186539e-010
1.27307840e-007   6.34206490e-013   3.52571891e-010   4.98335413e-016
5.48046529e-013   1.95679737e-017   3.45472611e-014   0.00000000e+000]
```

Figure 7: p-value of 64 features

of trees 20, depth of trees 4, max features 64. The average RMSE is 0.0297. After tuning the parameters, we have the best RMSE 0.00943. The parameters are : number of trees 32, depth of trees 12, max features 64. The best RMSE is shown in figure 8.

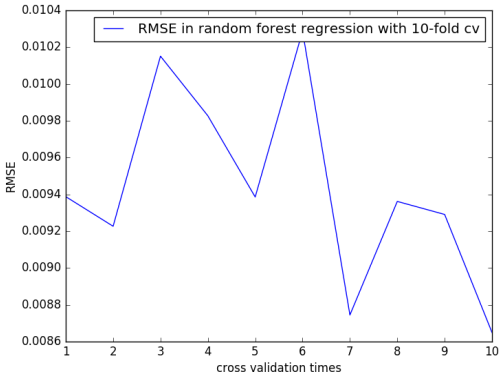

Figure 8: The best RMSE of random forest regression

The comparison between Linear Regression and Random Forest Regression is shown in figure 9. Random forest is a lot better than linear regression.

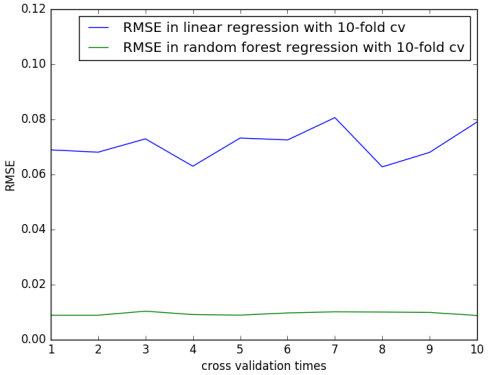

Figure 9: The comparison between Linear and Random Forest Regression

The prediction of the random forest model is shown in figure 10. We can see the cycle period is around 42 backup times, which is about a week, so the pattern is the same with the actual values.

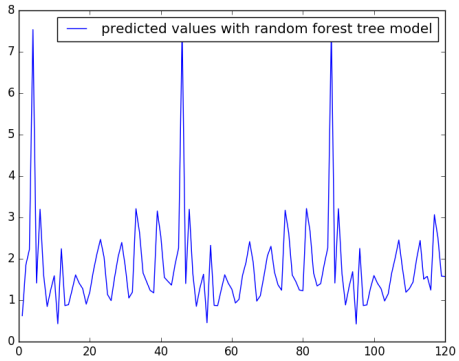

Figure 10: The prediction of the random forest

4 Neural Network

In the neural network model, we choose the feedforward network with the back-propagation trainer. The main parameters are the number of layers, the kind of hidden layers and the number of iteration before convergence. After lots of experience, we find that sigmoid layer is proper for hidden layers and linear layer is proper for the output layer. The optimal number of layers is 7. The RMSE reduces when the number of iteration is increasing, however it is too slow with large iteration number. So we set it to be 10. The best RMSE we can get is 0.0827. The RMSE is shown in figure 11.

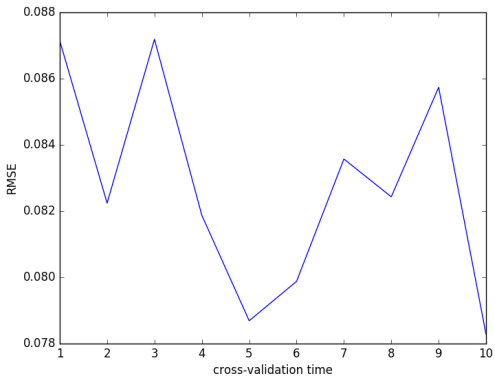

Figure 11: The RMSE of neural network model