Prediction of Bike Rental using Model Reuse Strategy

Arun Bala Subramaniyan and Rong Pan

School of Computing, Informatics, Decision Systems Engineering,
Arizona State University, Tempe, USA.
{bsarun, rong.pan}@asu.edu

Abstract This paper describes the methodology used for ECMLPKDD 2015 Discovery Challenge on Model Reuse with Bike Rental Station Data (MoReBikeS). The challenge was to predict the number of bikes in the new stations three hours in advance. Initially, the data for the first 25 new stations (station 201 to 225) was provided and various prediction methods were utilized on these test stations and the results were updated every week. Then the full test data for the remaining 50 stations (station 226 to 275) was given and the prediction was made using the best method obtained from the small test challenge. Several methods like Ordinary Least Squares, Poisson Regression, and Zero Inflated Poisson Regression were tried. But reusing the linear models learnt from the old stations (station 1 to 200) with lowest mean absolute error proved to be the simple and effective solution.

1 Introduction

Majority of the knowledge intensive application areas have a high chance of operating context variation. The reuse of the learnt knowledge might play a critical importance in generalizing the notion of the operating context. In this ECMLPKDD 2015 Discovery Challenge, the bike rental stations located in Valencia are considered. The objective is to predict the number of bikes available in each new stations (Station 201 to 275) three hours in advance. There are at least two use cases given for such predictions [1]. First, a user plans to rent (or return) a bike in 3 hour time and wants to choose a bike station which is not empty (or full). Second, the company wants to avoid situations where a station is empty or full and therefore needs to move bikes between stations. The data set consisted of all the necessary details like location, time, weather and profile of bike availability for model building and prediction.

2 Methodology

In order to make a successful prediction, the information about the current status in the station, the weather condition and the time period at which the stations would be empty or full were considered along with the profile of bike availability in each station which was learnt from the historical information. This is because the quality of the prediction can be the increased by collecting more historical information. Considering all the above given information, various methods like Ordinary Least Squares, Poisson Regression and Zero Inflated Poisson Regression were tried.

Apart from the above information, the linear models developed for old stations (station 1 to 200) based on the training dataset and their respective MAE values were available. After trying out various methods for prediction, the idea of reusing these models learnt from the old stations (station 1 to 200) to predict the number of bikes in the new stations (station 201 to 275) provided the best solution based on the MAE value. The selection of best models for the new stations and prediction of results is discussed in this section.

2.1 Model Extraction and Prediction

There were 7 base models available and in addition to that, 6 trained models were provided for each of the 200 old stations. As the deployment data for stations 201 to 275 was given, all the given linear models were utilized for predicting the number of bikes in each of the stations 201 to 275. The model with less Mean Absolute Error (MAE) was selected as the best model for a particular station. This process continued for selecting the best model for all the new stations (201 to 275).

In some cases, the prediction results were negative or it exceeded the maximum limit of the bikes that can be accommodated in a station. To overcome this problem, a constraint was added in such a way that whenever the result is negative, the predicted value is reset to zero and whenever the result exceeded the maximum limit, the value is reset to the number of docks at that station. So, this helped in reducing the MAE value further. Also, in some stations, the extraction algorithm came up with two or more models with the same MAE values. In those cases, only the first model was selected.

After the extraction algorithm selected the best models for each of the new stations based on the given criteria, the number of bikes in each station for the leaderboard data set were predicted using the extracted models. The same set of constraints were applied to avoid negative values and over fitting during prediction. The R software was used for model extraction and prediction. The MAE values for the small test challenge using this strategy was 2.502 and the MAE values for the full test challenge turned out to be 2.067.

3 Other Methods Tried For Prediction

Initially, before reusing the given linear models, new models were built with the deployment data for the stations 201 to 275. The different approaches used for building the models and their results are discussed in this section. The Minitab and R software is used for this purpose. As the test statistics and graphs for all stations cannot be included in this paper, a sample station data is chosen for illustration and understanding. Similar procedures were adopted in building models for all the other stations.

3.1 Ordinary Least Squares Method

After cleaning the given dataset, the first model was built using all the regressors under consideration. A thorough analysis of this full model, including residual analysis and multicollinearity check was done. Also, the scatter plot was used to study the relationship between the regressor and response variable. From the model summary, there was severe collinearity problem between the regressors. Also, the test statistics showed that only few variables significantly contributed to the model. The scatter plot of those variables is shown in Figure 1. The variable ‘y’ denotes the number of bikes. The variables x20, x21, x23 and x16 denotes bikes 3 hours ago, full profile bikes, short profile bikes and temperature respectively. The coefficient of determination value was not satisfactory and the PRESS (prediction sum of squares) statistic was large, making the model doubtful for prediction purposes.

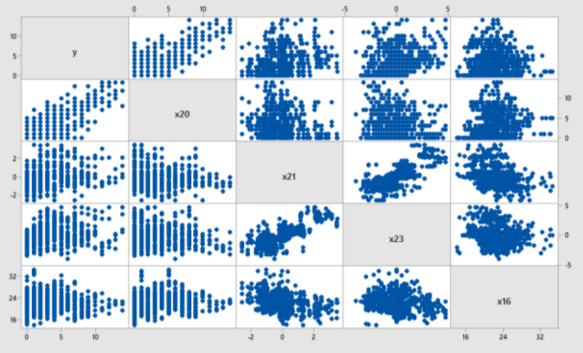

Fig.1. Scatter plot for the initial model

The residual plot for the initial full model is shown in Figure 2. The normal probability plot shows some deviations at the upper and lower tails. This might be due to the reason of existence of many zeroes in the response variable. The residual plot (deleted residuals versus the fitted values) shows a significant double bow pattern, violating the assumption of constant variance. Also, there are some outliers noticed from the residual versus observation plot.

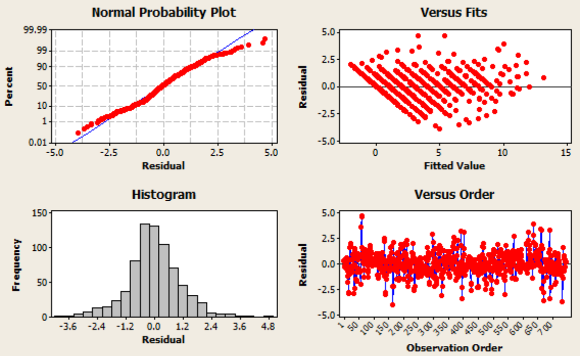

Fig.2. Residual plot of initial full model

To explore about the outliers, the values of ordinary residuals, studentized residuals, leverage (HI1), Cook’s distance, DFFIT were collected. Though some outliers were observed, no influential points were noticed which was confirmed from the cook’s distance. Since the reason for the unusual observations were not explicit, these observations were not removed and included for modelling.

In order to identify the regressors that were contributing to the model, the subset regression was done. The Mallows C_p and R-squared values were used in determining the best set of regressors. Care was taken to choose less number of regressors with low C_p and high R-squared value. Also, the stepwise regression, forward selection, backward elimination techniques were used. The alpha values for entering and removing the variables were set at 0.1 and 0.2 respectively. Finally, the regressors that significantly contributed to the model were identified.

After selecting the best subset of regressors, the analysis was carried out once again. The multicollinearity problem disappeared which was confirmed from the Variance Inflation Factor (VIF) values (less than 5). The PRESS statistic showed drastic improvement. Also, the significance of the regressors was examined.

The residual plot for subset regression is shown in Figure 3. Though the model improved slightly, there is a problem with normality assumption. The residual plot does not show any improvement as the double bow pattern still exists. This strongly suggested a need for variance stabilizing transformation of the variables along with the addition of polynomial and interaction terms for further improvement.

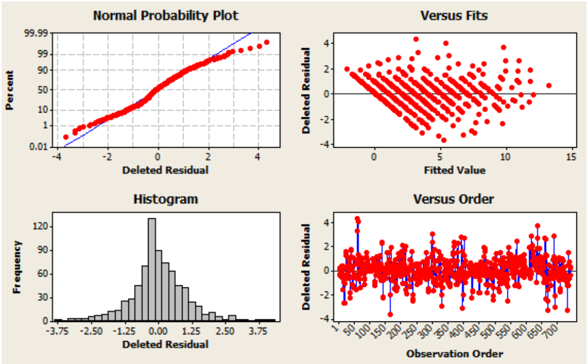

Fig.3. Residual plot of subset regression

All the possible sets of transformations (from square root to inverse) were tried on the response and regressor variables. Also, the models with polynomial terms and interaction terms were built. Finally, the logarithmic transformation of the regressor variables was tried and regressed against the response. This logarithmic transformation was a good choice for the model since the data involved historical information.

The ANOVA table provided all the necessary test statistics. The regressors that contributed significantly to the model were identified. There was an evidence of lack of fit for some models but it did not affect much. The PRESS statistic was low but the R-squared value dropped further. There were no traces of multicollinearity and the model seemed perfect.

The residual plot for final model is shown in Figure 4. The normal probability plot still needs some improvement but the variance is much stabilized. There is no pattern evident from the residual plot.

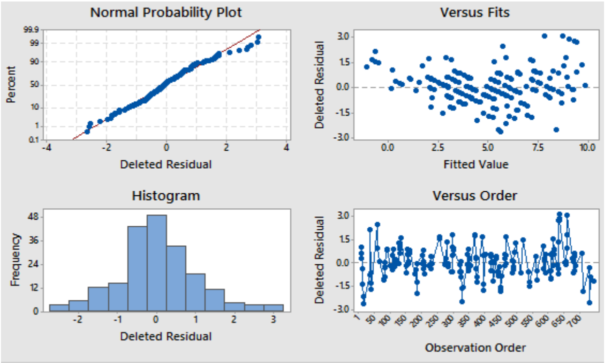

Fig.4. Residual plot of final model

3.2 Poisson Regression

In order to improve the model further and make it useful for prediction, the Poisson Regression was tried. The reason for choosing Poisson regression was that the response variable involved counting the number of bikes, which was discrete. The log link was particularly attractive for Poisson regression as it ensured that all of the predicted values of the response variable will be nonnegative.

The initial full model was fitted with Poisson regression. This model seemed to be good when compared to the final model built using the ordinary least square method. There were some regressors which were not significant, noticed after examining the test statistic and also their regression coefficients were negligible.

The Poisson regression along with the stepwise selection of regressors was done in order to obtain the best subset of regressors. The final set of regressors seemed to be almost the same as in case of ordinary least squares method. The test statistic summary was used to understand the significance of regressors. The R-squared value improved slightly for this initial model. The Akaike Information Criteria (AIC) was also high, which denoted the expected entropy of the model was maximum. The key insight provided by the AIC value is similar to R-squared adjust and Mallows C_p. The multicollinearity problem was studied from the VIF values. The standard residuals, studentized residuals, cook's distance, leverage values were examined

The Goodness of fit test provided the value of deviance with its significance. The ratio of deviance to the degree of freedom value was near to unity. The Pearson chi squared test value was also small with larger p-value indicating that the fit was significant. Also, the partial deviance test indicated that all the selected regressors were significant to the model.

The residual plot for the Poisson regression is shown in Figure 5. The upper tail of the normal probability plot seems to be good but there is some problem with the lower tail. Also, the assumption of constant variance is violated, observed from the plot of deleted residuals versus fitted values. There is a nonlinear pattern observed in this plot indicating a need for transformation and higher order terms.

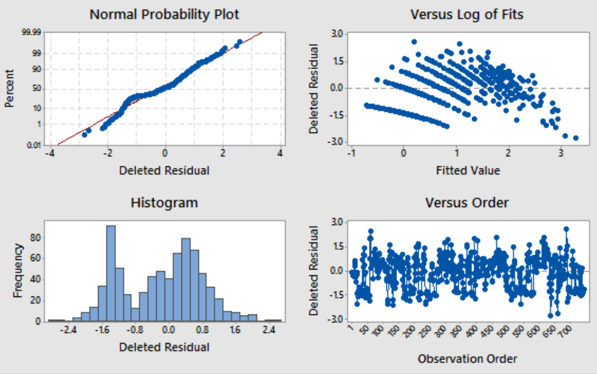

Fig.5. Residual plot for initial model of Poisson Regression

Various transformations were tried out and the final combination of variables was found. The natural log link function was used and the logarithmic transformation of the regressors proved to be good. All the test statistics were examined once again. Finally a better model when compared to all the previous models was obtained.

The deviance table provided all the necessary test statistic. There were no traces of lack of fit. The error values were low and no traces of multicollinearity was observed from VIF values. The Confidence Interval limits were shrunken, which was good. The R-squared value was good and the model seemed to be perfect. The residual plot for the transformed model is shown in Figure 6.

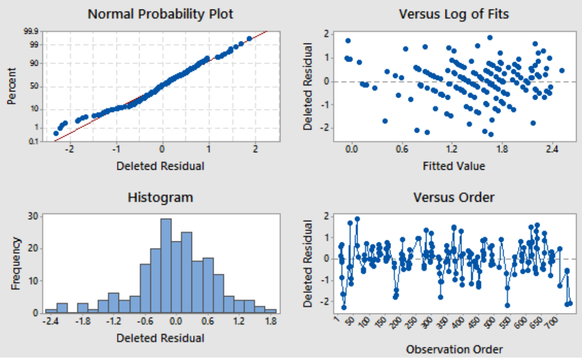

Fig.6. Residual plot after transformation

The upper tail of the normal probability plot is good but there is still a problem at the lower tail. But the assumption of constant variance is satisfied as observed from the plot of deleted residuals versus fitted values. The values of the residuals are distributed within a range of 4 (+2 to -2). There is no pattern observed from the plot and the model has improved a lot when compared to the previous models.

Only thing that troubled much is the lower tail of the normal probability plot. The presence of excess zeroes in the response than usual observations could have resulted in larger residuals in the prediction. The existence of these excess zeroes also caused trouble in fitting the model. So, in order to overcome this problem, Zero Inflated Poisson Regression was tried.

3.3 Zero Inflated Poisson Regression

As there were excess of zeroes when examining the response data, there arose a doubt that some of the zeroes might be inflated. So, in order to solve this problem, the Zero Inflated Poisson Regression was tried. The glm2, ggplot and pscl packages were used for zero inflated poisson regression in R software. Finally two models were generated, one for the count model and other for the inflated zeroes.

The best subset of regressors were selected and the model was built and analyzed. The pearson residual was low. The R-squared value was similar to poisson regression and also prediction sum of squares statistic was small relative to the other methods. Apart from that, the log likelihood values were large enough with good significance, indicating that one or more of the regressors in the subset contributed significantly to the model. There was no evidence of lack of fit and multicollinearity. The count model seemed to fit the data well. From the zero inflated model, the various factors that contributed towards inflation of zeroes were identified.

The normal probability plot (Q-Q plot) and the residual plot was studied. The normal probability plot improved further when compared to the previous methods. The residual plot did not have any problem apart from some outliers as shown in Figure 7.

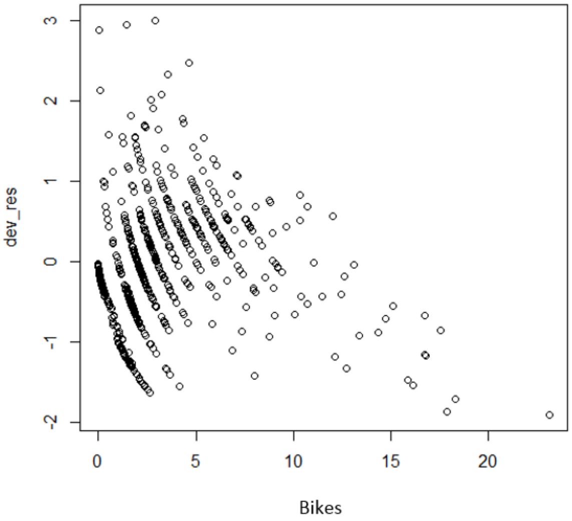

Fig.7. Deviance Residual plot of initial model

To improve the model further, transformation of the variables was done and the results of the transformed model was studied. The results obtained after the transformation and addition of interaction terms improved the model further. All the test statistics similar to the poisson regression model were checked. The normal probability plot and the residual plot is shown in Figures 8 and 9 respectively. The Zero Inflated Poisson model had only less number of terms and found to fit the given data well. The validation of regression models is discussed in the next section.

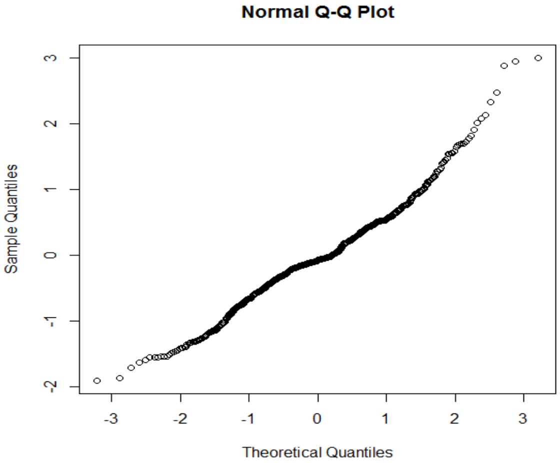

Fig.8. Normal probability plot of the transformed model

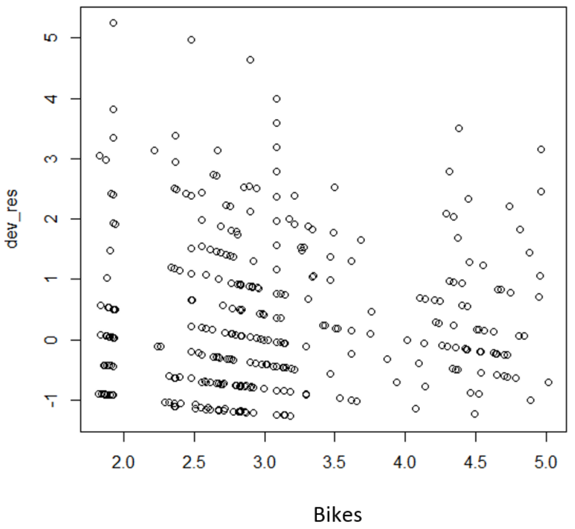

Fig.9. Deviance Residual plot of the transformed model

4 Validation of Regression Models

After the final model is built, it has to be validated to check whether the model is adequate for prediction. Model validation is directed towards determining if the model will function successfully in its intended operating environment.

Initially the new models were built based on the deployment data for the month of October 2014. As the data for the next month was not available, data splitting technique was used for validation. But the prediction capability of the model for November 2014 was still doubtful by this method of validation. Also, the results of the small test challenge were not satisfactory.

So, the validation approach was modified. As the training dataset for the stations 1 to 10 were provided, the above mentioned model building approaches were tried on the training dataset for October 2013 and MAE values were calculated by predicting the bikes for November 2013. This method of validation seemed to be a good approach and it revealed some interesting facts. The model without transformation and addition of interaction terms had low MAE values when compared to a model with many terms. The model with large number of terms fitted the given data well but in case of prediction it was overfitting the data. Also, the leaderboard results of the small test data supported this claim. The MAE values for the prediction using models from OLS method, Poisson Regression and Zero Inflated Poisson Regression were 2.724, 3.068 and 2.774 respectively. The MAE value for the models with transformation and interaction terms was larger than the baseline value of 3.288. So the models built by transforming the regressors and adding interaction terms did not work well for predicting the bikes in this challenge.

Even though these methods worked well, their MAE values were still larger than the MAE values obtained from reusing the old models, which was 2.502. So, reusing the models seemed to provide better results as they were obtained from the training data sets of the stations. So, this method was selected to predict the number of bikes in the full test data.

5 Results and Discussions

Finally, the idea of reusing the linear models built from the old stations was better than building new models for the given deployment data. This was obvious because, the old models were obtained from the training dataset with data collected over two years, but the deployment data was just for a month.

Though R-squared values increased after transforming the regressors and including interaction terms, the model was not suitable for prediction, which can be confirmed from the MAE values of small test challenge. As the number of terms increased, there was a risk of overfitting. The model with simpler terms worked well for this challenge. Also, the models built using the training data predicted the results better than the newly built models with limited data. This indicated that the models should be robust in order to account for variations in the data. Even though the models were built for some other stations, they seem to predict well for new stations than the models built using only the deployment data of new stations. Also, a good validation approach should be used for choosing the best models.

One more approach that seemed to work was modelling of error values. This was carried out in order to reduce further variation in the selected model. This was done by collecting the error values from fitting each new station data by reusing the models from old stations. These error values were treated as response variable and regressed against the new station variables to build a model. Now, the model selected from the old stations along with the model created from the error values were combined to form a new model for the station. In addition to this, Lasso Regression was tried but these methods were not included for predicting the full test set in this challenge. Also, rounding the values affected the MAE values. In most cases, the MAE values decreased after rounding the results. But, for some cases, rounding the values did not have much effect.

Thus, it is understood that the reuse of the learnt knowledge can play a critical importance in generalizing the notion of the operating context.

References

1. REFRAME (2012 – 2016), ECML/PKDD Discovery Challenge #1, 2015, "*MoReBikeS: Model Reuse with Bike rental Station data*". http://reframe-d2k.org/Challenge

2. Montgomery, D., Peck, E., & Vining, G. (2012). *Introduction to linear regression analysis* (5th ed.). Hoboken, NJ: Wiley.

3. Cameron, A., & Trivedi, P. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics, 46*, 347-364.

4. Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics, 34*(1), 1-14.

5. Allison, P., & Waterman, R. (2002). Fixed-Effects Negative Binomial Regression Models. *Sociological Methodology, 32*, 247-265.

6. Hans, C. (2009). Bayesian lasso regression. *Biometrika, 96*(4), 835-845.

7. Zou, G. (2004). A Modified Poisson Regression Approach to Prospective Studies with Binary Data. *American Journal of Epidemiology, 159*(7), 702-706.