STAT 685: Dr. Suojin Wang's Group

Modeling Seoul Bike Sharing Demand

Nam Tran, Bai Zou

November 3, 2020

Contents

List of Figures

List of Tables

Chapter 1

Dependency of Y_t

1.1 Autocorrelation and Partial Autocorrelation of Rented Biked Count, y_t

In section 2.2, the ACF and PACF plots are suggesting strong autocorrelation of the past bike demand. The ACF shows a clear daily trend (every 24 lags). The hour feature used in estimators can be used to reflect this daily trend. The PACF shows significant dependence between demand and past demands - lag 1, 2, 3, 4, 5, 8, 9 ,10, etc. However, this past demand information is not well used in previous estimators. A simple way to use the information is to add dependency features of past demand.

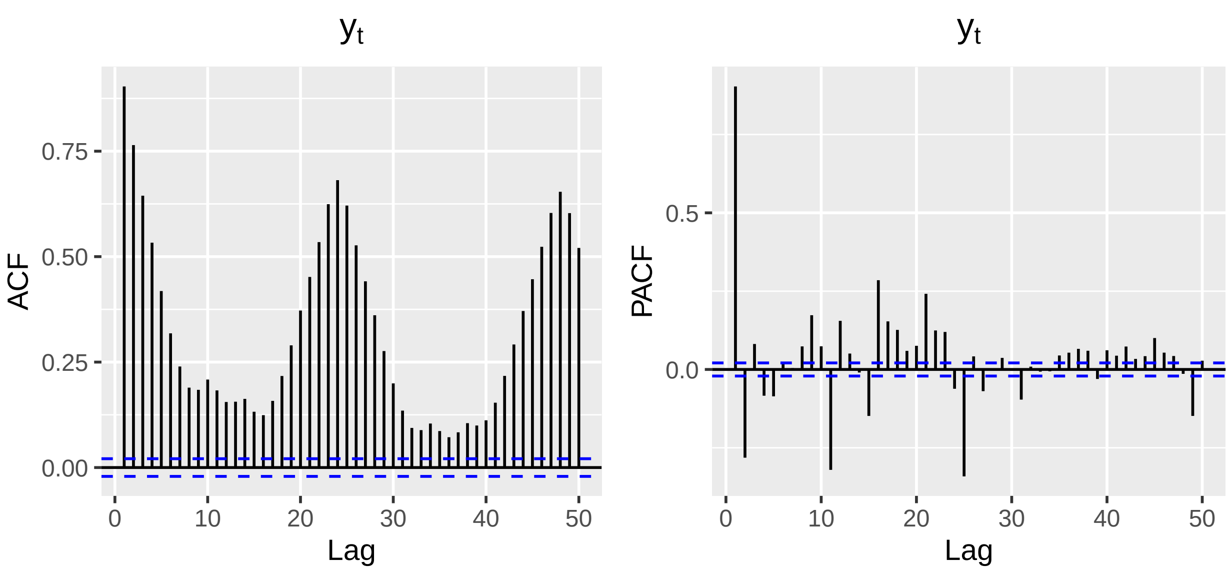

Figure 1.1: Autocorrelation and Partial Autocorrelation of Rented Bike Count, i.e., y_t

1.2 Dependency Feature

In addition to weather feature and hour feature, add dependency features for past demand: e.g., lag_1 means demand in the last hour; lag_4 means hourly demand 4 hours ago, etc. By adding dependency features, the top observations will have missing value as past demand information is not available. The obaservation with missing values will be deleted from training data.

1.2.1 Estimation without Dependency Features

In previous study, boosting method brings the best result among all estimation methods. Therefore, the estimator using boosting method with parameter tuned from previous study will be used in all following model training and predictions. All observations before anchor date Nov 1, 2018 will are used for training and the 30 observations from Nov 1, 2018 to Nov 30, 2018 are used for testing. To compare the impact of dependency feature, model M_0 is fit with only weather features and hour information. The test R^2 in M_0 is 0.756. The figure below is showing the feature importance from M_0, which suggesting 'Temp' as the most important features effecting the demand.

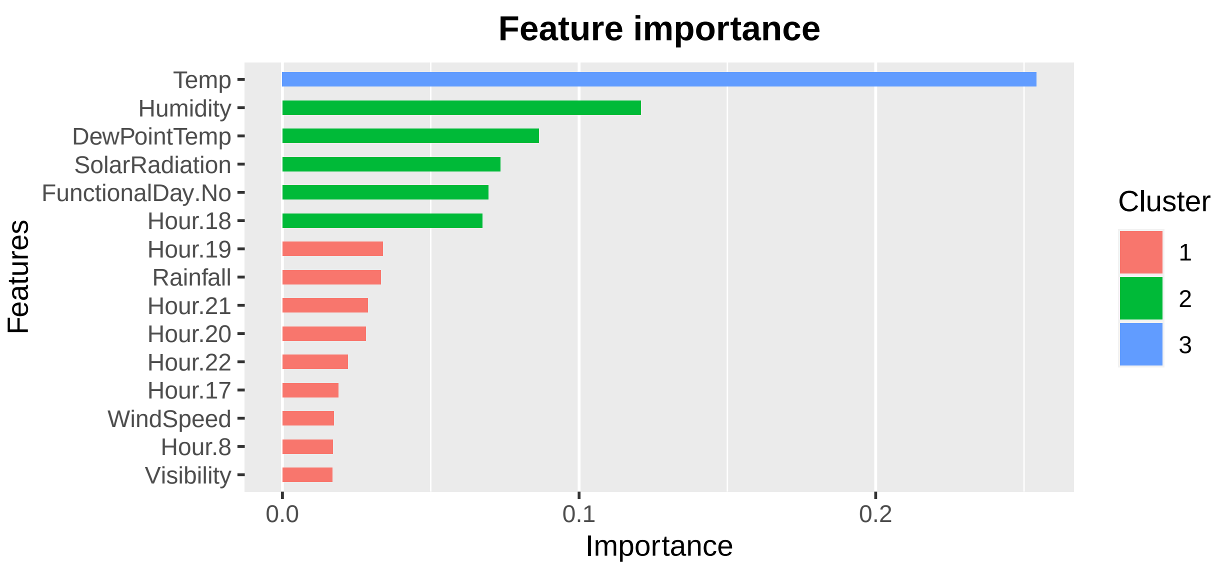

Figure 1.2: Feature Importance M_0

1.2.2 Estimation with Dependency Features

Based on PACF plot, dependency features within 1 day (24 hours) are selected for the model fitting, including the past 1-24 hour demand except for lag 6, 7, 14, as M_1. With the additional 21 columns added to X, the test R^2 goes up to 0.967 - a significant improvement from M_0. In addition, the feature importance plot for M_1 below shows the past demand features are more important than weather features compared with M_0.

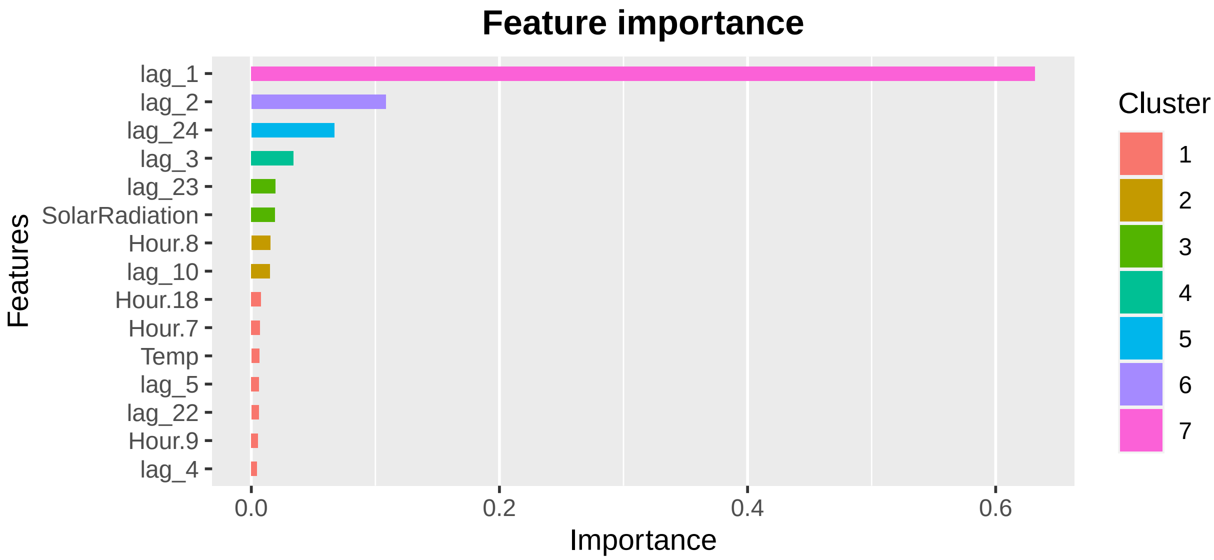

Figure 1.3: Feature Importance M_1

1.2.3 Estimation with Reduced Dependency Features

Based on the feature importance plot of M_1 and the PACF plot, the dependency features can be reset to lag 1, 2, 11, 16 and 24 to reduce the size of X variables. Thus M_2 is fitted with smaller number of dependency features and the test R^2 is 0.958, which is less than 1% lower than M_1 and significant higher than M_0.

1.2.4 Estimation with Dependency Features Defined by Business Assumptions

In M_3, the dependency features added is lag 1, 2, 24, 48 and 24×7 based on business assumptions that the demand is related to the past demands 1 and 2 hours ago, 1 day ago and 1 week ago. The test R^2 is 0.95, which is as good as M_1 and M_2.

1.2.5 Estimation Comparison

The table below is comparing the test R^2 among models with different dependency features. Considering the test accuracy and variable size, M_2 and M_3 are preferred.

Table 1.1: Test R2 Comparison of Models with Different Dependency Features

	Model	TestR2
No_Dependency_Feature	M0	0.756
Full_Dependency_Feature	M1	0.967
Reduced_Dependency_Feature	M2	0.958
Business_Dependency_Feature	M3	0.950

Chapter 2

Forecasting Application

2.1 Business Scenaros

The purpose to predict bike demand is to make bikes available and accessible to the public at the right time. Thus, the forecasting of hourly bike demand is required to support business decisions and operations. Based on system infrastructure capacity, we define two typical business scenarios below in real application.

2.1.1 Daily Data Update

The system data is updated on a daily base for further analysis and forecasting to the next 24 hours' **hourly demand** is required for high-level planning of the next day. To test the performance in this business scenario, we assume the data is updated at 0:00 AM of the day and all available data at the moment is used for model training to predict the next 24 hour demands.

2.1.2 Real-time Data Update

If the system infrastructure could support real-time data update, an hourly model training could be run to predict the next **hour demand**. Any changes in the past hours could be used in the next hour demand prediction. To test the performance, the models will be trained hourly with all the data available at the moment (include demand data from last hour) and used to predict the demand in the next coming hour.

2.2 Hourly Demand Forecasting with Daily Data Update

2.2.1 Estimator and Dependency Features

When data is updated once every day, the latest observation available to use as dependency feature is the lag 24 for all prediction time stamps. Therefore, the smallest lag of dependency features can be added is lag 24. Based on the business assumption in dependency feature study, lag 24, 48 and 24×7 are added as dependency features, which represent the past demand from same hour 1 day ago, 2 days ago and 1 week ago. The model training is repeated daily for November 2018. And the forecasting method is boosting using the parameters tuned from previous study.

2.2.2 Forecasting Results

In a daily data update, there are 30 repeated model training and forecasting (1 in each day). In each iteration of model training and forecasting, all observations prior to the iterator date are used as training data and the next 24 hours' hourly demands are used as testing data. The training R^2, mean CV R^2 and testing R^2 results are recorded in each iteration.

The table below (see Table @ref(tab:daily_tab)) shows the average training R^2 over the 30 iterations is 97.7% and the average mean CV R^2 over the 30 iterations is 79.9%. The low average value and large std value in testing R^2 represents some poor prediction accuracy in some of the iterations. This can tell clearly in the figure below comparing forecasted demand and real demand (See Figure @ref(fig:daily_fig)) - at Date Nov 3, Nov 6 and Nov 9, when there's no demand due to non-functional day, the forecasting at the beginning of the day still predicts certain amount of demand.

The overall forecasting R^2 is calculated between all forecasted demand and real demand of the 30 iterations and will be used to compare forecasting accuracy in all scenarios.

Table 2.1: Foresting Result with Daily Data Update and Dependency

	Average	Std
Train_R2 (per train)	0.977	0.001
Mean_CV_R2 (per train)	0.799	0.003
Test_R2 (per train)	-0.156	4.544
Overall_Forecasting_R2	0.860	NA

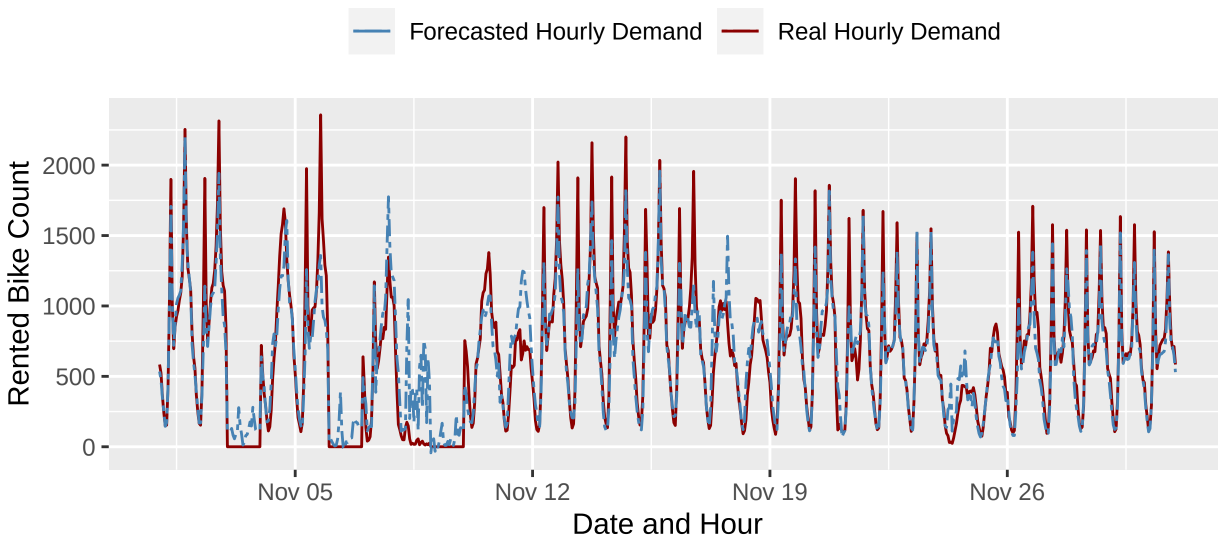

Figure 2.1: Forecasted Demand and Real Demand Comparison with Daily Data Update and Dependency

2.3 Hourly Demand Forecasting with Real-time Data Update

2.3.1 Dependency Features

When the data is updated in real-time, all past demands up to one hour ago (lag 1) can be used as dependency feature. Therefore, we add the past demand from 1 hour ago, 2 hours ago, 1 day ago and 1 week ago as dependency features to the data (M_3 in previous study). The modeling training is repeated every hour and used to predict demand only for the coming hour. And the forecasting method is boosting using the parameters tuned from previous study.

2.3.2 Forecasting Results

In the real-time data update, there are 30×24 repeated model training and forecasting (1 in each hour). In each iteration of model training and forecasting, all observations prior to the iterator date and hour are used as training data and the next 1 hour demands is used as testing data. As there is only 1 observation in the testing data, testing R^2 is not available.

The table below (see Table @ref(tab:hourly_tab)) shows both average training R^2 and average mean CV R^2 over the 30×24 iterations are over 90%. The overall forecasting R^2 reaches 96%. And the figure comparing forecasted demand and real demand below (See Figure

@ref(fig:hourly_fig)) shows a very good match between the two curves, representing an accurate prediction. Moreover, with the real-time data update, the system is able to know the latest demand in the past hour and adjust the coming hour demand prediction - forecasted demand in Nov 3, Nov 6 and Nov 9 stays low when detecting low demand in the last hour.

Table 2.2: Foresting Result with Real-time Data Update and Dependency

	Average	Std
Train_R2 (per train)	0.995	0.000
Mean_CV_R2 (per train)	0.929	0.001
Test_R2 (per train)	NA	NA
Overall_Forecasting_R2	0.964	NA

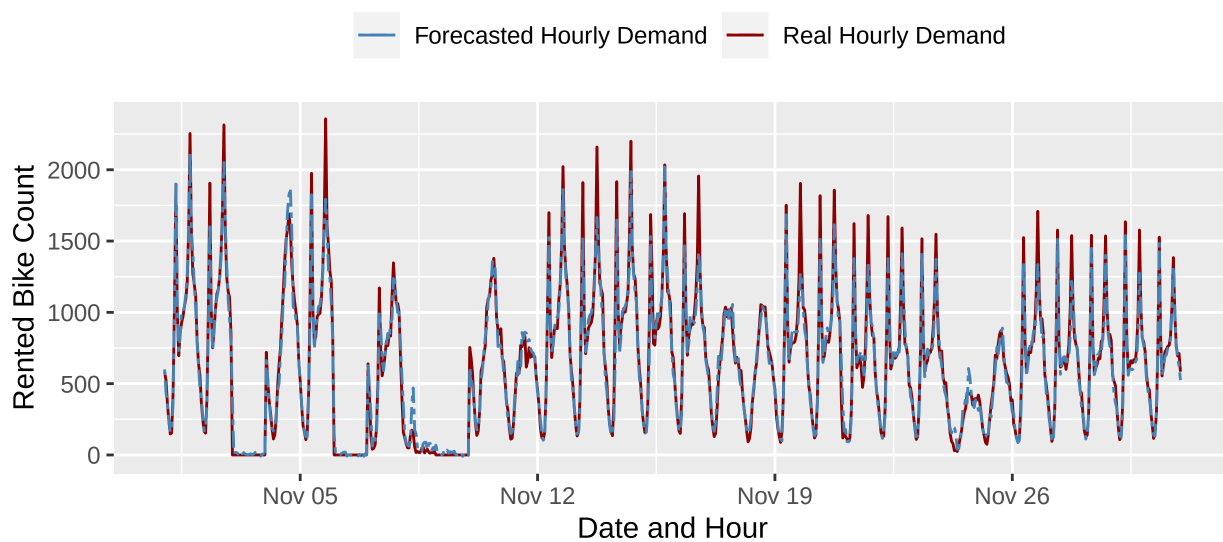

Figure 2.2: Forecasted Demand and Real Demand Comparison with Real-time Data Update and Dependency

2.4 Forecasting Results Comparison

2.4.1 Improvement with Real-time Data Update

If the system could support real-time data update, the overall forecasting R^2 shows a 10% improvement (see Table @ref(tab:compare_tab)). Comparing the forecasted demand to real demand, the real-time data update forecasting shows a much lower discrepancies than the daily data update forecasting (See Figure @ref(fig:compare_fig) and @ref(fig:compare_fig1)).

Table 2.3: Forecasting Results Comparison with Dependency

	Daily Data Update with Dependency	Real-time Data Update with Dependency
Train_R2 (per train)	0.977	0.995
Mean_CV_R2 (per train)	0.799	0.929
Test_R2 (per train)	-0.156	NA
Overall_Forecasting_R2	0.860	0.964

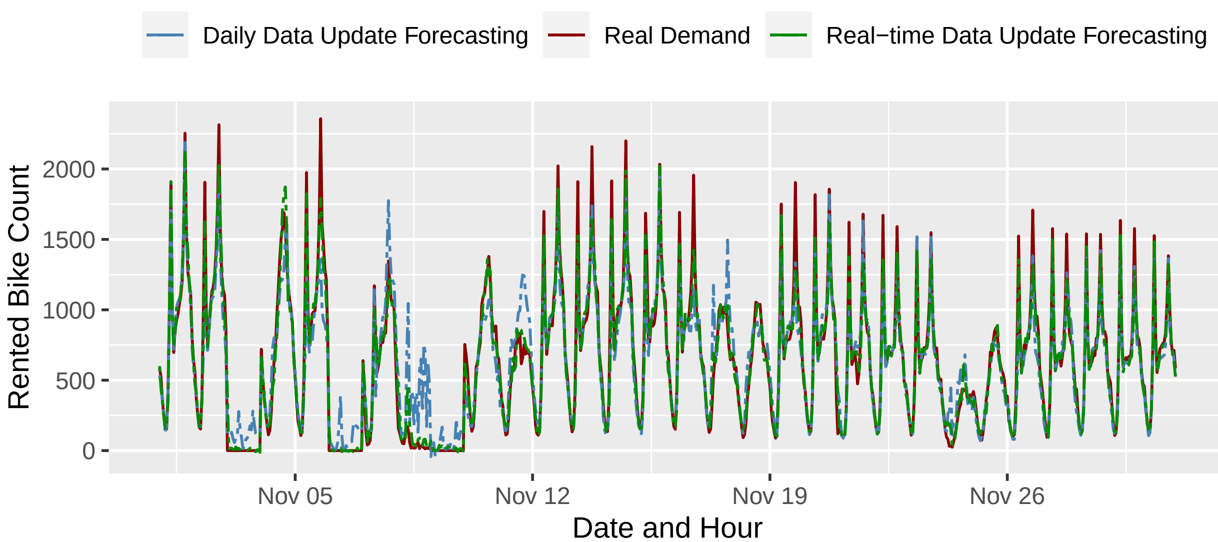

Figure 2.3: Forecasted Demand Comparison with Dependency

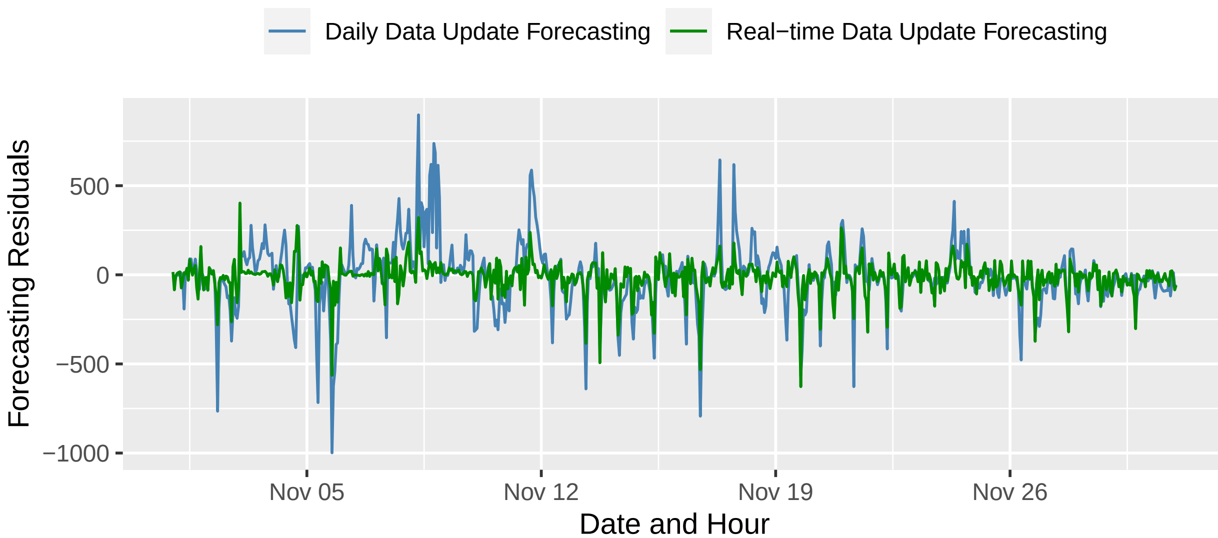

Figure 2.4: Forecasting Residual Comparison with Dependency

2.4.2 Improvement with Dependency

To better understand the importance of dependency, the same forecasting studies (repeated model training and forecasting on daily base and hourly base) are conducted without dependency features.

The table below (see Table @ref(tab:compare_tab2)) shows 8.5% improvement in daily data update and 14.7% improvement in real-time data update in terms of overall forecasting R^2. Comparing the two figures plotting forecasting residuals (See Figure @ref(fig:compare_fig3) and @ref(fig:compare_fig4)), there's more significant improvement by adding dependency features with real-time data update than the scenario of daily data update - the blue residual line in the second figure is much more smooth than the green line compared with the first figure.

Moreover, if the system could not support a real-time data update, using the dependency features in the daily data update scenario still brings better (4.3% higher) forecasting accuracy than a real-time data update without dependency.

Table 2.4: Forecasting Results Comparison with and without Dependency

	Daily Data Update with Dependency	Real-time Data Update with Dependency	Daily Data Update without Dependency	Real-time Data Update without Dependency
Train_R2 (per train)	0.977	0.995	0.960	0.960

	Daily Data Update with Dependency	Real-time Data Update with Dependency	Daily Data Update without Dependency	Real-time Data Update without Dependency
Mean_CV_R2 (per train)	0.799	0.929	0.686	0.686
Test_R2 (per train)	-0.156	NA	-0.219	NA
Overall_Forecasting_R2	0.860	0.964	0.775	0.817

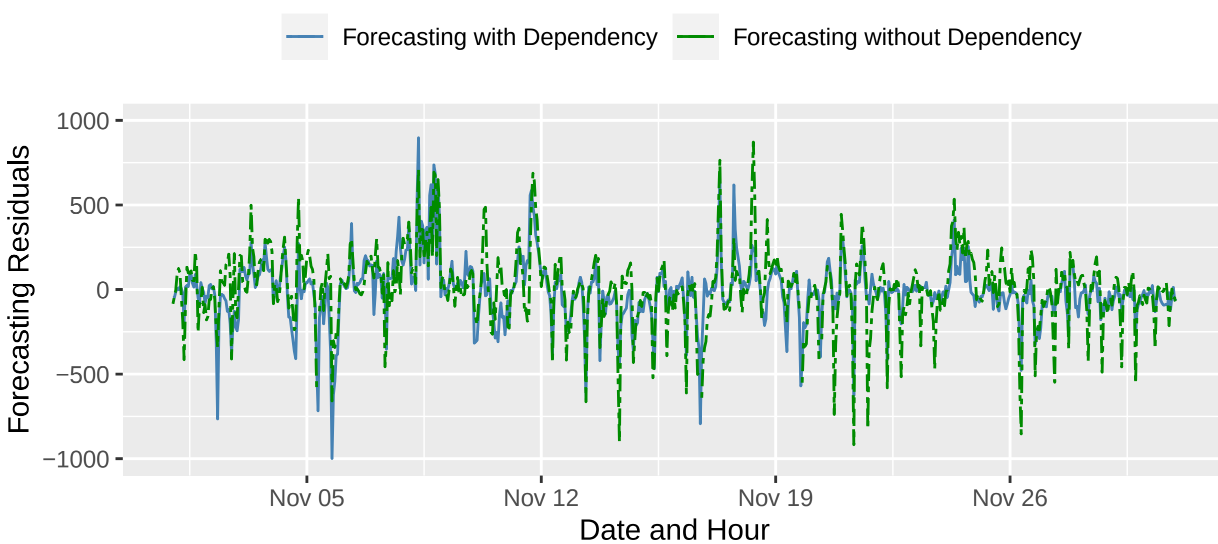

Figure 2.5: Forecasted Demand Comparison with Daily Data Update

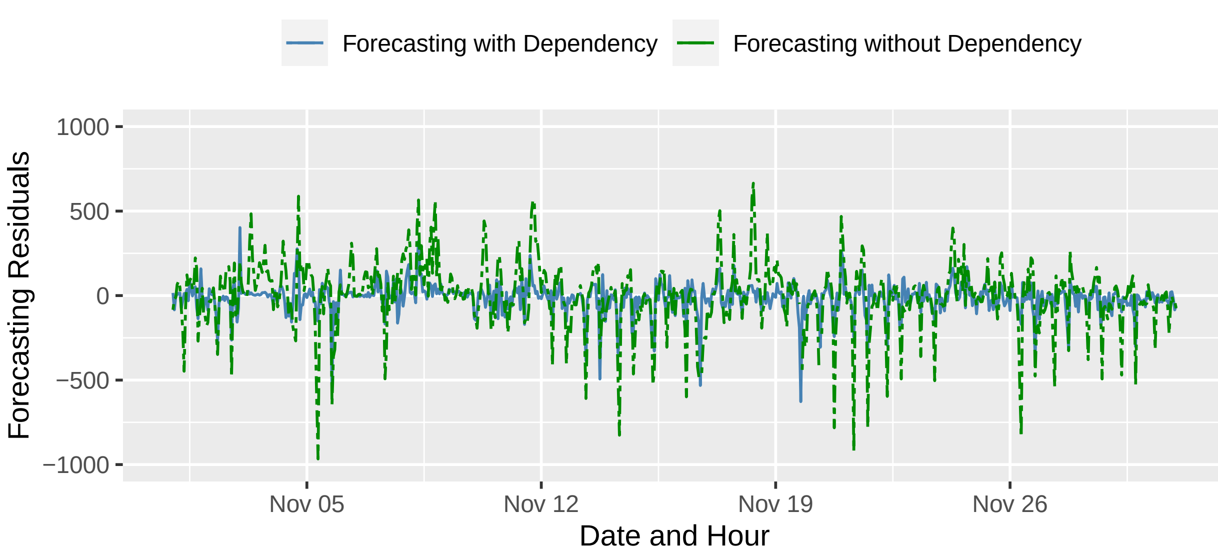

Figure 2.6: Forecasted Demand Comparison with Real-time Data Update