Popular science summary
Time series prediction for algorithmic rescaling in the cloud

Björn Elmers
Supervisor: Kalle Åström

February 4, 2016

DICE is a Swedish software company that develops computer games. To make their games run smooth at all times they have to make sure that their computers always have enough server capacity. The old way of doing this is to try to guess how much server capacity that is needed when there is most players online and then buy that many computers. This means that during the morning, when not very many players are gaming, a lot of their server capacity will be unused. The new way of doing this is with the cloud. What I mean by that is that instead of buying their own machines they can rent server capacity from other companies such as Amazon or Google that already have lots of computers. This is a lot more expensive than to own the computer yourself but the trick is that you can buy and sell on-demand. So you can rent more capacity during the "rush-hours" in the evening and not as much during the slow morning hours.

The thing is, that buying cloud space and starting up a new game server takes 5-15 minutes which means that you need a forecast, or prediction, of how many players you will have 5-15 minutes in the future. And how do you get such a prediction? Well, you use historical data, like how many players did we have at this time yesterday. Then you use that information to say something about the future. For instance you could say, "I guess that there is as many players 13.00 today as it was 13.00 yesterday".

In my thesis I have tested different types of forecasting techniques, or prediction models as I call them. The model that made the most accurate guesses was an auto regressive model of order 100. What it does is that it

takes parts of old values up to 100 steps back and adds them together. So maybe it takes 1.7 times the last value, 0.1 of the second last value and 0.1 of the value 30 steps back and adds them together. You can see in the figure below that the predictions mostly lands very close to the real values. The curves almost melts together.

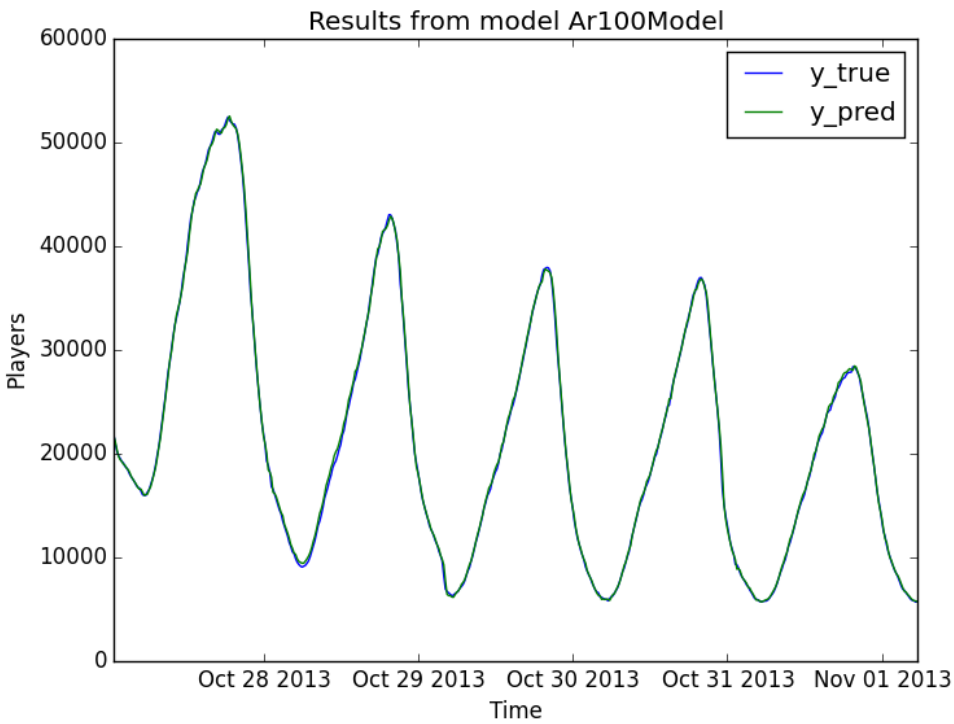

Figure 1: The true values together with the prediction.

Another thing that took a lot of my focus was how to clean my data before the prediction. Almost all data from real life is dirty in some way. My data contained a lot of outliers. That is values that has been influenced by something other than what you want to look at. In this case it was influenced by problem with the servers and I did not want to use them in my predictions. The tricky part was to find out if a value was an outlier or not and to solve this I developed an algorithm. The algorithm simply said that if there was too much difference between the last value and the current one then the current one was flagged as an outlier.

Yet another thing I tried to accomplish with my paper was to help increase the general understanding of the player behaviour. One of the things I found was that there is a very strong weekly season in the dataset. Another thing was that the player behaviour almost looks identical Mondays to Thursdays while Fridays and Saturdays had a later decrease in players than the other days and Saturdays and Sundays had an earlier increase. This can be seen in the figure below where first each day is plotted separately and then the mean of each weekday is plotted separately.

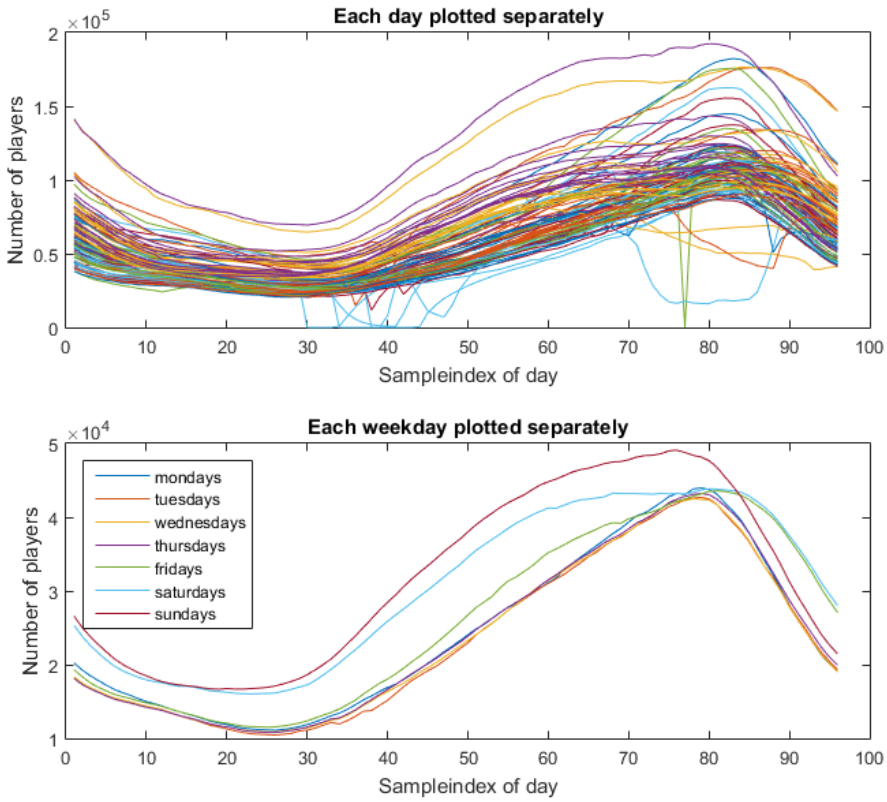

Figure 2: First each day in the dataset is plotted separately then the mean of each weekday is plotted.