Marc Lange
Tel.: 0171 - 110 39 21
eMail: marc@lange-iz.de

Hamburg, 21. Oktober 2016

**Analysis of the Bike-Sharing Dataset**

The dataset has the form

| instant | dteday | season | yr | mnth | hr | holiday | weekday | workingday | weathersit | temp | atemp | hum | windspeed | casual | registered | cnt |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2011-01-01 | 1 | 0 | 1 | 0 | 0 | 6 | 0 | 1 | 0.24 | 0.2879 | 0.81 | 0 | 3 | 13 | 16 |
| 2 | 2011-01-01 | 1 | 0 | 1 | 1 | 0 | 6 | 0 | 1 | 0.22 | 0.2727 | 0.8 | 0 | 8 | 32 | 40 |
| 3 | 2011-01-01 | 1 | 0 | 1 | 2 | 0 | 6 | 0 | 1 | 0.22 | 0.2727 | 0.8 | 0 | 5 | 27 | 32. |

where *instant* is the key for the row, *dteday* is a date-string for the event, *season* is the seasons spring, summer, autumn, winter encoded from 1 to 4, *yr* encodes in 0/1 if the event happened in 2011/2012, *mnth* and *hr* directly encode month and hour. The column *weekday* encodes weekdays from 0 to 6 starting with Sunday in the US-American way, *workingday* encodes if the date was a regular working day, while *holiday* notes holidays. The following columns each describe the weather: *weathersit* describes the badness of the weather in numbers of 1 to 4, *temp* and *atemp* describe the actual and the apparent temperature (according to an unidentified index for this). Furthermore *hum* describes humidity and *windspeed* describes wind speed. Finally the last three columns *casual, registered, cnt* describe the users in that timeframe that rented a car, casual the unregistered users, registered the registered ones and *cnt* is there sum. The datasets *day.csv, hour.csv* only differ in that the users are aggregated by either days or hours (hence with additional column *hr* as described above).

My task was to analyse which variables influence the customer's decision to rent a bike, specifically to predict 'cnt' given the other data, which we can each collect in advance, or in case of the weather at least get good predictions for.

To that end I fit a linear regression model onto each dataset, and found that by backward elimination we get the following 0.95-significant variables:

day *yr, mnth, workingday, weathersit, temp, windspeed, Saturday,*
hour *yr, mnth, hr, holiday, workingday, weathersit, temp, windspeed, Saturday,*

where *Saturday* is a derived variable which yes/no-marks the Saturdays in our dataset. Using 10% of each dataset as a test-dataset, the regression produce mean average errors of

day ca 700,
hour ca 110,

i.e., on average we predict the daily user count with an error of 700 people, while we can predict hourly users with an error of about 100 people.

## The Models

The analysis of daily users results in the following linear regression model (by backward elimination with $\alpha = 0.05$ with coefficients rounded to integers).

| const | yr | mnth | holiday | workingday | weathersit | temp | windspeed | Saturday |
|---|---|---|---|---|---|---|---|---|
| 1513 | 2072 | 89 | -508 | 307 | -762 | 5472 | -2784 | 345. |

The analysis of hourly users with the same method gives the following model:

| const | yr | mnth | hr | holiday | workingday | weathersit | temp | windspeed | Saturday |
|---|---|---|---|---|---|---|---|---|---|
| -121 | 88 | 4 | 9 | -20 | 12 | 29 | 304 | 106 | 14. |

The apparent increase in hourly users with windspeed and weather situation has to be an error in the hourly model, caused by the collinearity of year and month as well as the fact that hours are obviously not a linearly contributing factor. In fact, plotting regression errors against hour, we find the most extreme errors at hours 7, 8, 17, 18, while the other errors are actually a good image of homoscedasticity. I.e., commuters are quite a non-linear contribution to our dataset, and to improve the hourly predictions if needed, I would first introduce an indicator for commuter's times, and convert yr + mnth into a decimal yrmnth = yr + mnth*1/12.

I shall disregard the hourly model now, since it is too error-ridden; the mean average error of about 100 already indicated that, where the daily model was at 700, so much smaller than $24 * 100 = 2400$.

## Interpreting the Daily Model Data

The daily model gives us the following insights: First the variables *yr, mnth* tell us this data observed a general trend of ever more increasing usage of this bike rental service. In a next pass I would want to remove this trend from the data and then analyse what apart from the general trend tells us the major influences on bike usage. However, we see that from year 2011 to 2012 we had an increase of about 2000 users per day, and a daily user increase of about 90 per month.

Apart from the time trend: Unsurprisingly temperature is a dominating factor – the increase between 0°C and 41°C is one of about 5500 users per day, which is drastic, while windspeed obviously drops daily usage by about 2500 users. Furthermore in the daily model we have a decrease of users by about 800 for each weather category.

Finally the influences in the hundreds are holiday, workingday and Saturday: Saturdays increase the number of users by about 350, while general holidays drop users by about 500, and working days increase by 300. In particular this

indicates heavily that this rental service is used most by people commuting by bike to and from work, since holidays hurt the user numbers while working days increase them, and Saturday seems to work against this trend. Maybe inquiring if the renters have jobs on Saturday would provide more insight.

### The Conclusion

We can obviously not control the weather variable, however, it indicates days when big overhauls of the bike inventory can be performed easily, since a difference in about 150 daily users per degree Celsius is actionable. The analogous insight does not work well for the days. The major decrease is on holidays, where the rental shop is supposed to be closed, too.

However, the general insight into the data is that the number of users is driven by the workforce. So to balance usage out, one would have to expand into the spare-time sector. The occasional advertisement of a biking tour in the area on a variety of channels would probably help, if sponsored by this rental shop. If the strategy were instead to actually focus on the workforce, then one could probably improve the service and thus the acceptance of the service by identifying, which are the hubs, where people would want to have their bike, and maybe transport them accordingly.

## Other Ideas

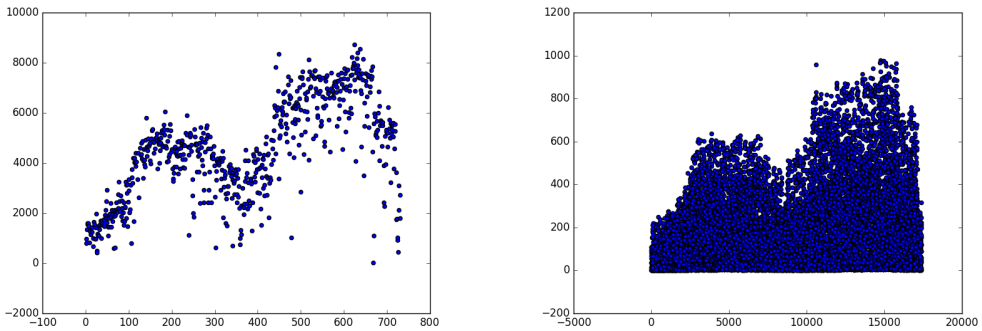
I started looking at the data visually too late, the simple plots

```
plt.plot(dfday.instant,dfday.cnt); plt.show
```

and

```
plt.plot(dfhour.instant,dfhour.cnt); plt.show
```

of key versus count produce quite the insightful pictures:




showing that there is indeed this downward parabolic trend with respect to the

seasons and the hourly data is just too smeared out by the presumably as well downward parabolic trend in the hours. In particular, we can see quite clearly that the linear regression is at most an adequate first pass for the daily data, but I would try to fit a model of $y \sim \boldsymbol{x} + sin(c \cdot \boldsymbol{x})$, where in hours we have sufficient data to even learn the coefficients $c$ modulating the period of the sine-function in the $x_i$'s. Presumably we would get a model with $c_i \neq 0$ for season and hour, thus compensating for the wavy form of the data.

On the other hand I would love to try the basic ensemble learning / boosting process of splitting the prediction error $\varepsilon$ of the linear regression on the test set into $\varepsilon \mapsto |\varepsilon| + sign(\varepsilon)$. Training a decision tree to predict $sign(\varepsilon)$ would give valuable insight into whether we over- or underestimate respectively and by inspecting the first branchings, what variables dominate that behaviour.