Weather and Bike Rentals

What are the influening factors and

how well can they predict the number of rented bikes?

Ingo Nader

Nov 2018

Abstract

This piece of work investigates **to what extent** and **how weather and time of day influence bike rentals** in a public bike sharing system in Montreal. Public data obtained from the Canadian government's past weather and climate service, as well as bike sharing data available from Kaggle are analyzed via a simple baseline model (moving average) and a more complex machine learning model (gradient boosting regression). Partial dependence plots (PDP) and individual conditional expectation plots (ICE) are used to visualize the influences of the different factors.

Results show that the model can explain the number of hourly bike rides very well (93.3% of variance explained). The most important influences on the number of bike rides seem to be temperature, atmospheric pressure, hour of the day and relative humidity, but there are strong interactions between these influences: For example, the number of predicted bike rides increases with temperature, but only if relative humidity is not too high.

The full analysis is available in multiple python files on github: kgl-cycle-share-main-file.py.
A synopsis is available as an ipython notebook cycle-share-analysis-synopsis.ipynb, or as html to download.

Motivation

There are quite a few analyses that **investigate bike sharing in relation to weather**, and also some freely available and easily retrievable datasets. But they often use daily aggregates (e.g., the UCI machine learning dataset on bike sharing), or they only have a relatively small sample size (e.g., Kaggle bike sharing demand). As I use my bike on a daily basis for (parts of) my commute, I found these investigations very interesting. However, I always thought that daily aggregate data is not enough to understand the phenomenon. Personally, I don't decide on the weather of the whole day if I'll be using the bike. Rather, I just decide if I can use the bike right now.

Hence, I wanted to do a **similar analysis myself, but on a more fine-grained level**, i.e., on hourly data, with a big enough sample size to try out some machine learning models.

Dataset(s)

Two datasets were used: **Bike sharing data**…

- **BIXI Montreal public bicycle sharing system**, North America's first large-scale bike sharing system
- Available via Kaggle from https://www.kaggle.com/aubertsigouin/biximtl/home
- For years 2014 to 2017
- Contains **individual records of bike trips**: timestamp and station code for start and end of trip, duration
- $n = 14598961$ records (individual bike trips)
- Station codes, names, and position (latitude, longitude) available in separate files, but only of secondary interest for this analysis

…and **weather data** from the Canadian government:

- Canadian government's past weather and climate service, available from http://climate.weather.gc.ca/ historical_data/ search_historic_data_e.html
- API for **bulk data download**: http://climate.weather.gc.ca/climate_data/ bulk_data_e.html
- Data can be downloaded per weather station per month and contains **hourly measurements** of different metrics (e.g., timestamp, temperature, relative humidity, atmospheric pressure, wind speed; different measures available for different stations)
- $n = 35064$ hourly weather records in total (between 672 and 744 per monthly file)

Data Preparation and Cleaning

- First, **data download** was performed manually for the bike share data from Kaggle (as only available after login), and via a Python script for the weather data (bulk download).
- For the weather data, the **weather station** that was most central to the locations of the bike rides was picked (see data exploration).
- Next, the **data was loaded** and contatenated into a pandas `DataFrame` each for individual bike rides and hourly weather data.
- The next step was **calculating the variable of interest: Hourly bike rides**. This was done by aggregating individual bike trips to hourly counts of trips (number of trips in each hour), using the starting time of the trip.
- Then, the **weather data was joined to the hourly bike ride data**, using the common timestamp as join key.
- One feature **(wind chill) was dropped**, as it had too many missing values (77.9% missing).
- Finally, **additional features were added** for the analysis: hour of the day (0-23), day of the week (0-6, zero corresponding to Monday, six corresponding to Sunday), month (1-12).
- These features, despite being categorical in nature, were kept as **continuous features**, as this proved to have more predictive power in the models.
- For modeling, **rows with missing values were dropped**, as the goal is not having the most complete prediction coverage, but rather an indication of the prediction quality that is possible with complete data. In total, 1284 rows (0.04%) of the original data were dropped.
- The remaining rows were **split into training and testing set** (90% of the data, $n = 26168$ rows for training, the remaining 10%, $n = 2908$ for testing).

Research Question(s)

The research questions that I wanted to answer with my analysis were:

- **To what extent do** the number of **bike rides depend on the current weather conditions**? That is, how well can the number of bike rides be predicted from weather data (and time of year, time of day)?
- What are the **most important factors** that influence the number of bike rides?
- **How do these factors influence** the number of bike rides? What are the main effects of these factors, and what are the interactions between them?

Methods

First, some **data exploration** was performed, in order to get to know the data and to find out how the number of hourly bike trips is distributed across the investigated time span. Also, the interrelation of features was investigated by means of a correlation heatmap.

In order to find out how well the number of bike rides can be predicted from the data, different models were used. As a **baseline model**, a moving average was calculated to find out how this very simple model can explain the data.

Then, after splitting the data into 90% training and 10% test set, **different machine learning models** were fitted to the data in order to predict the hourly number of bike rides from the available data: Random forest regression, and gradient boosting regression via `scikit-learn` and `xgboost`. The most promising model, `scikit-learn`'s gradient boosting regression, was fitted via a randomized 4-fold cross-validation for indentifying the best hyperparameters. Variable importance was used to identify the most important influence factors, and *partial dependence plots* (*PDP*) [1] and *individual conditional expectation* (*ICE*) plots [2] were used to **visualize the influences of the important variables** on the number of bike trips.

Findings: Data Exploration

The Canadian government's **past weather and climate service** provides data from weather stations around the whole country. Fortunately, it offers a *search by proximity* function. Via this service, some sample data of the closest stations to Montreal were downloaded. Each of the data files contains information about the weather stations, including the geographical position (latitude and longitude). These coordinates were plotted on a map (see Figure on the right), and the **closest station** to the bulk of the data was chosen (station name: *MCTAVISH*). Only data from this station was used.

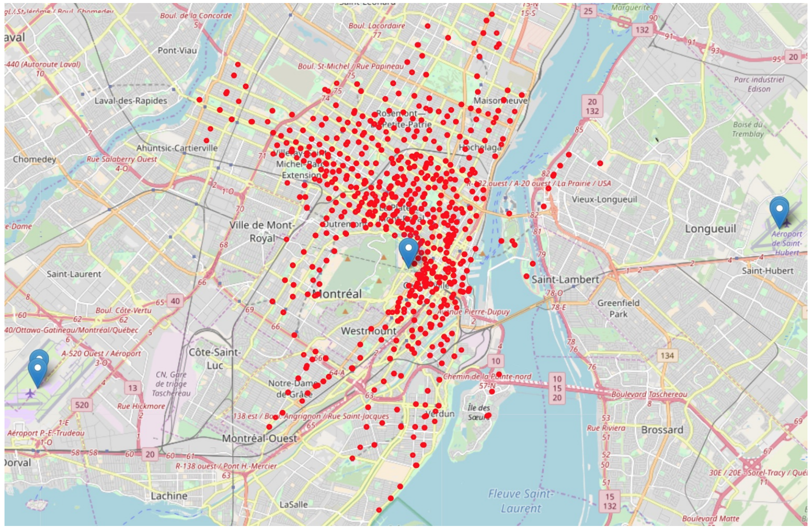


Figure: Plot shows all starting stations of a bike trip (red dots), as well as the closest weather stations (blue markers). The closest station in the center is the *MCTAVISH* weather station (Climate Identifier 7024745, WMO Identifier 71612)

Findings: Data Exploration

To get a better understanding of the data, the **number of hourly bike trips was visualized** for the time span between 2014 and 2017.

The **moving average** that is shown in the plot (red line) can be interpreted as a **baseline model**, i.e., the simplest possible model to describe the hourly number of bike rides.

This baseline model explains 38.8% of the variance $(r^2 = 0.388)$ and has a mean absolute error of $MAE = 316.2$, which means that on average, the prediction for the number of hourly bike rides is wrong by this many bike rides. This includes also winter months with no rides. For a more realistic estimation of model quality, these numbers drop to $r^2 = .079$ and $MAE = 510.7$ when only considering the time frame from May to September.

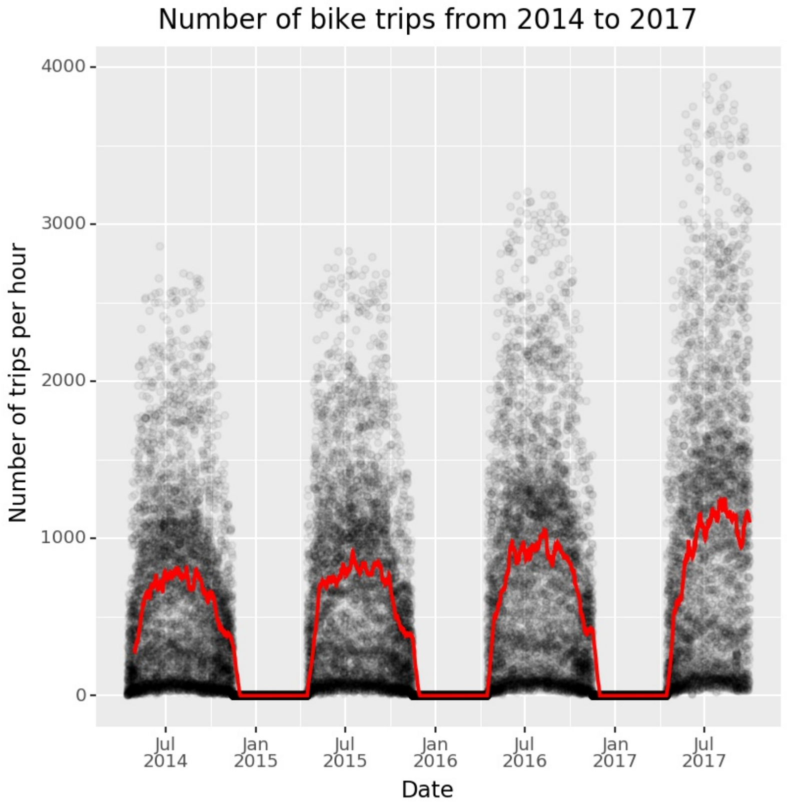


Figure: Number of hourly rides from 2014 to 2017. Each dot represents the number of trips in one specifc hour. Red line represents a moving average using a window of 14 days.

Findings: Data Exploration

To visualize the relations between the available features, a **correlation heatmap** is shown on the right. The features are only slightly correlated, with the only exception being temperature and dew point that show an almost perfect (linear) relationship $(r = .93)$.

To avoid problems resulting from this multicollinearity, only temperature was used as a predictor, and dew point was dropped. Despite the fact that that gradient boosting is less influenced by multicollinearity, it might still influence calculations of variable importance [3,4].

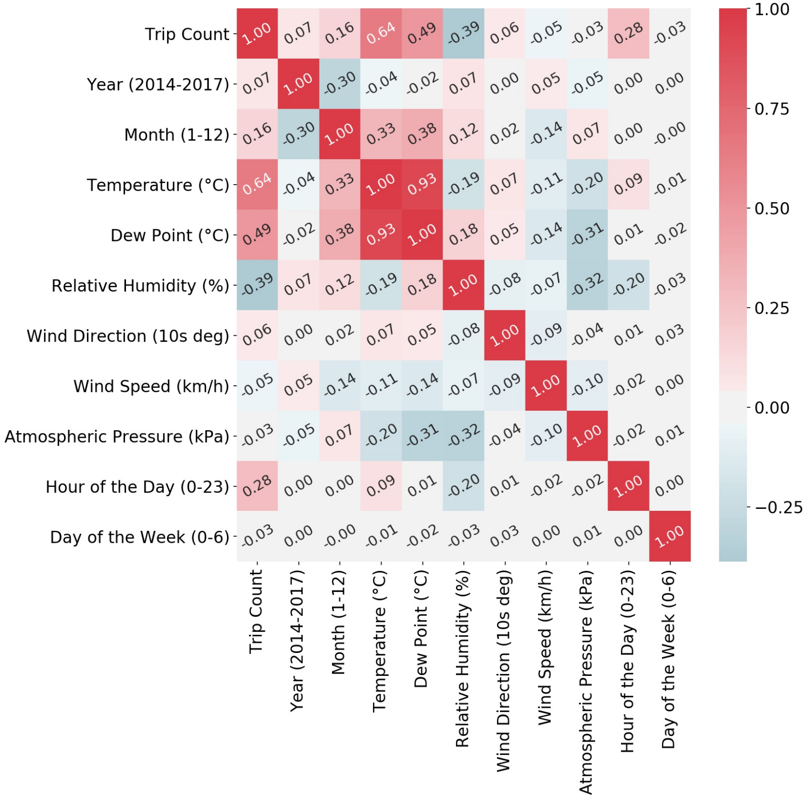


Figure: Pearson Correlations between available features in the data.

Findings: Model Fit

The explained variance of the different models tried ranged from $r^2_{test} = 0.865$ to $r^2_{test} = 0.941$ for the final model (all in the test set; see table). The hyperparameters for the final model (gradient boosting via scikit-learn) were selected via randomized search using 4-fold cross validation (using only the training set).

The **final model fits the data very well**, explaining 94.1% of the variance and exhibiting an average error of 85.4 rides over all hourly predictions. This error increases only slightly (to 105.4) when only summer months are considered.

Table: Model performance measures for different models. r^2 is the amount of variance explained, MAE stands for mean absolute error. $train$ and $test$ specify training and testing set, respectively. For the test set, some performance measures were also re-computed for using only the summer months of the test set (May to September), indicated via the $summer$ subscripts.

Model	r^2_{train}	r^2_{test}	r^2_{summer}	MAE_{test}	MAE_{summer}
Gradient Boosting (XGBoost)	0.896	0.865	0.856	155.3	173.1
Random Forest	0.913	0.894	0.880	111.2	139.9
Gradient Boosting (sklearn)	0.997	0.941	0.933	85.4	105.4

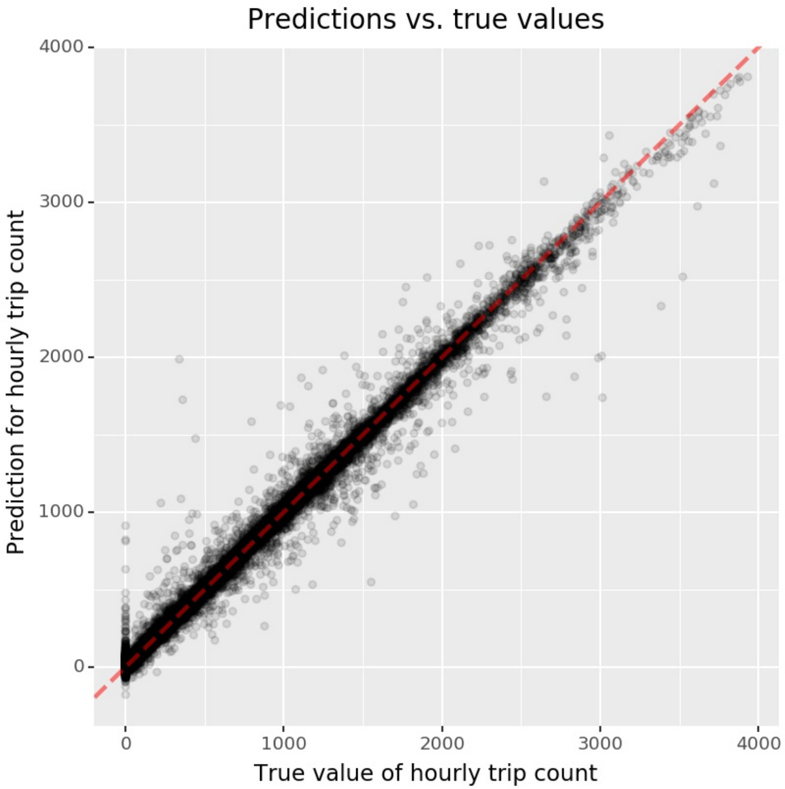


Figure: The figure shows actual (true) values vs. predicted values for the number of hourly bike trips for the final model. A perfect model would yield predictions that are identical to the true values, i.e., all points would be on the 45° diagonal. This model is relatively close.

Findings: Most Important Features

The **most important features** that influence the prediciton of hourly bike rides were **temperature** and **atmospheric pressure**. While the former is easily comprehensible, the latter is probably best understood as a proxy for precipitation, which was not available from the weather data: low pressure is commonly related to rainy weather [5]. While **relative humidity** is not as tightly connected to rain [6], it interacts with temperature, e.g., it influences how (high) temperatures are perceived [7].

Further important predictors are the **hour of the day** and the **day of the week**, as well as **wind direction and speed**. How these features influence the predicted number of bike trips is best detailed by specific plots, so-called *partial dependence plots* (*PDP*) [1] and *individual conditional expectation* (*ICE*) plots [2].

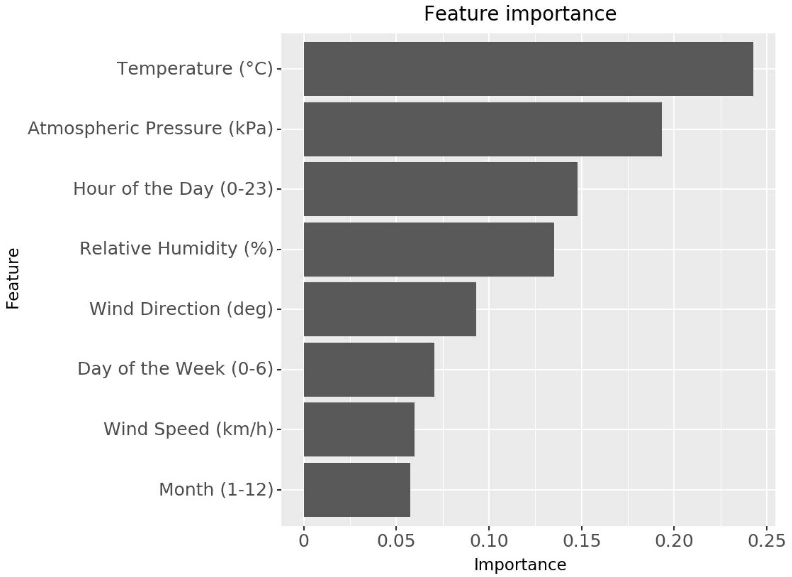


Figure: Most important features in the best-fitting model (gradient boosting via scikit-learn). The plot shows the most important features and their relative importance for predicting the hourly number of bike trips. The importance value indicates how well a feature performs in order to reduce the mean squared error for predicting the outcome, averaged over all trees in the ensemble. Values are normed and cannot be interpreted on an absolute scale [8].

Findings: Main effects

Regarding **hour of the day**, rush hours are clearly associated with higher number of hourly bike rides. There is a peak at eight o'clock in the morning and a (less pronounced) one in the afternoon. It is also visible that hour of the day interacts with other variables. First, the average PDP line is misleading, as there is a dense area of parallel lines indicating higher bike rides, as well as a high density at zero bike rides (i.e., the average PDP line obscures these two clusters). Second, There are a few lines showing a different pattern, starting to pick up only after the rush hour and steadily rising until the afternoon. This might be the interaction with the day of the week (see section *Interactions*; see also figure caption for more details about PDP and ICE plots).

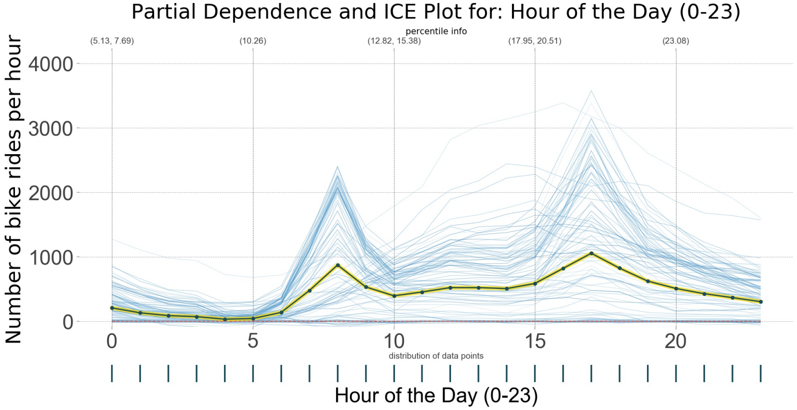


Figure: ***Partial dependence plots*** (*PDP*) [1,2] show the influence of a feature over its whole range (x-axis). The predicted outcome is visualized on the y-axis. That is, the **thick, dark blue line** in the plot shows the main effect of a feature. In addition, ***individual conditional expectation*** (*ICE*) plots [2] quantify the extent of interactions that exist with other variables. The **fine, light blue lines** in the plot show predictions at different levels of the feature in question. If they these lines are parallel, there are no interactions with other features. If they are not and show varying slopes compared to the main affect, the feature interacts with other features.

In addition, the plot shows the **distribution of the data** over the whole feature range as dark blue ticks below the y-axis. For hour of the day, this is not very interesting (only full hours), but it shows areas with sparse data for other features.

Findings: Main effects

Bike rides also increase with the **temperature** (the most important feature), once it has reached a certain treshold of about two or three degrees celcius. The increase is relatively steady up to a temperature or about 25 degrees celcius, where the number of hourly rides starts to level off and then to slightly decline. It is also visible (from the spread of the ICE plot lines) that there are massive interactions of temperature with other features.

For **relative humidity**, the main effect is less pronounced. The general tendency is a slight decrease for higher humidity, with a more pronounced drop around 95%. But the main effect is relatively weak compared to the overall spread of the ICE plot lines, and there are strong interactions with other features for higher relative humidity.

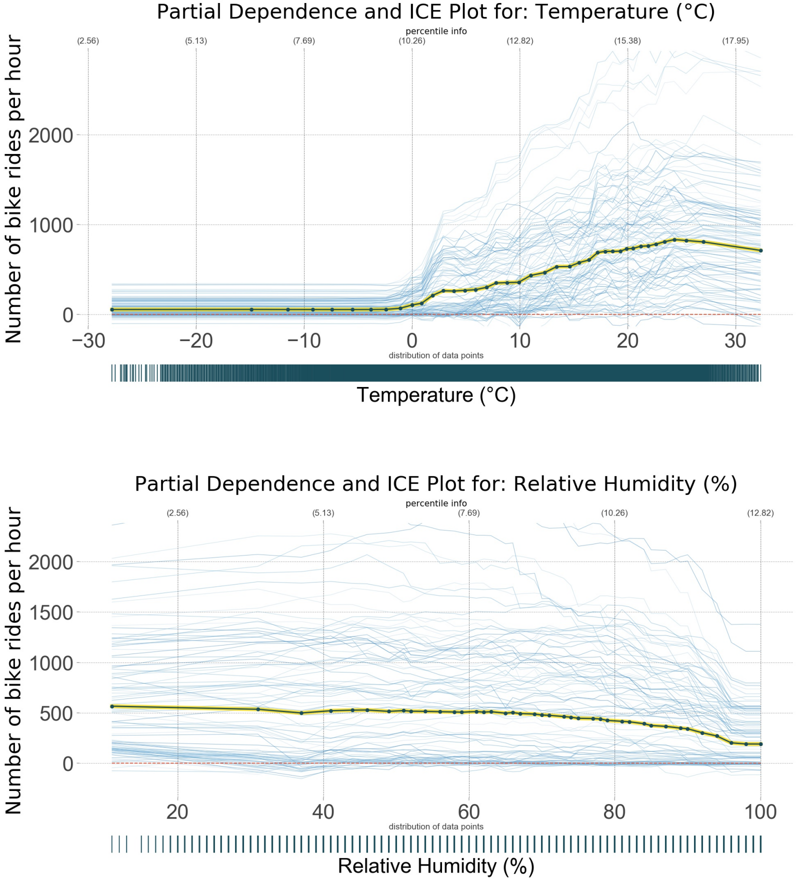


Figure: *Partial dependence plots* (*PDP*; thick, dark blue lines) [1] and *individual conditional expectation* (*ICE*) plots [2] (thin, light blue lines). For a more detailed explanation, see first occurence of this plot type on the previous page.

Findings: Interactions

The interaction between **day of the week and hour of the day** is not the strongest interaction in this model (see note on the side), but it serves well to describe the two-feature partial dependence plots (for details, see figure caption). It is clearly visible that during week days (0-4), there are peaks in the number of predicted hourly bike rides at 8 o'clock in the morning, as well as an even more pronounced peak in the afternoon (higher predicted numbers, as well as a broader peak). The number or predicted bike rides is slightly lower on Monday mornings.

During the weekend (days 5 and 6), the distribution is different (i.e., there is an interaction between the two features). There are by far fewer predicted bike rides, also the peak is in the afternoon. Additionally, there are more bike rides during the night, as compared to weekdays.

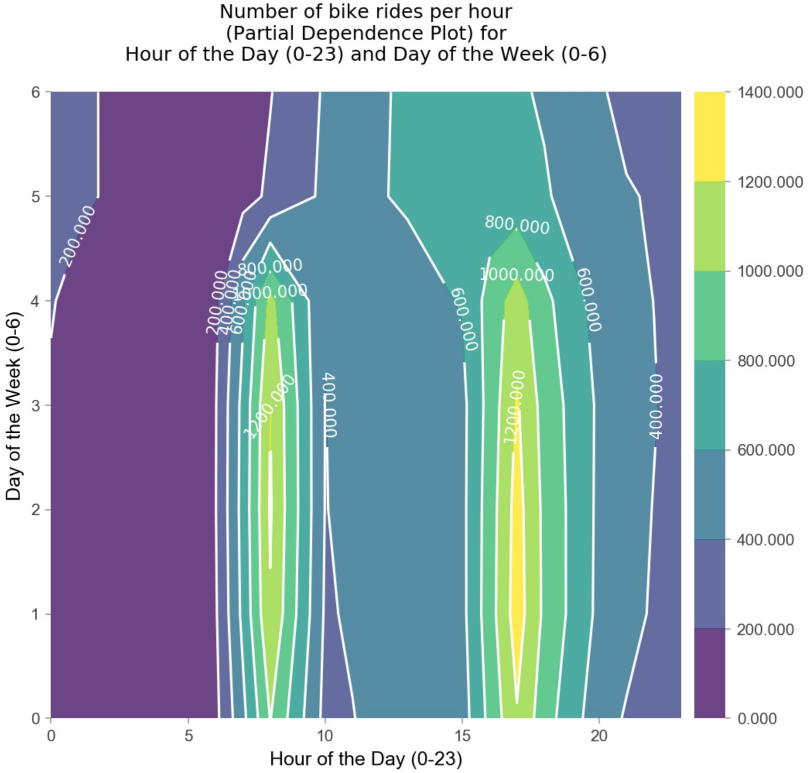
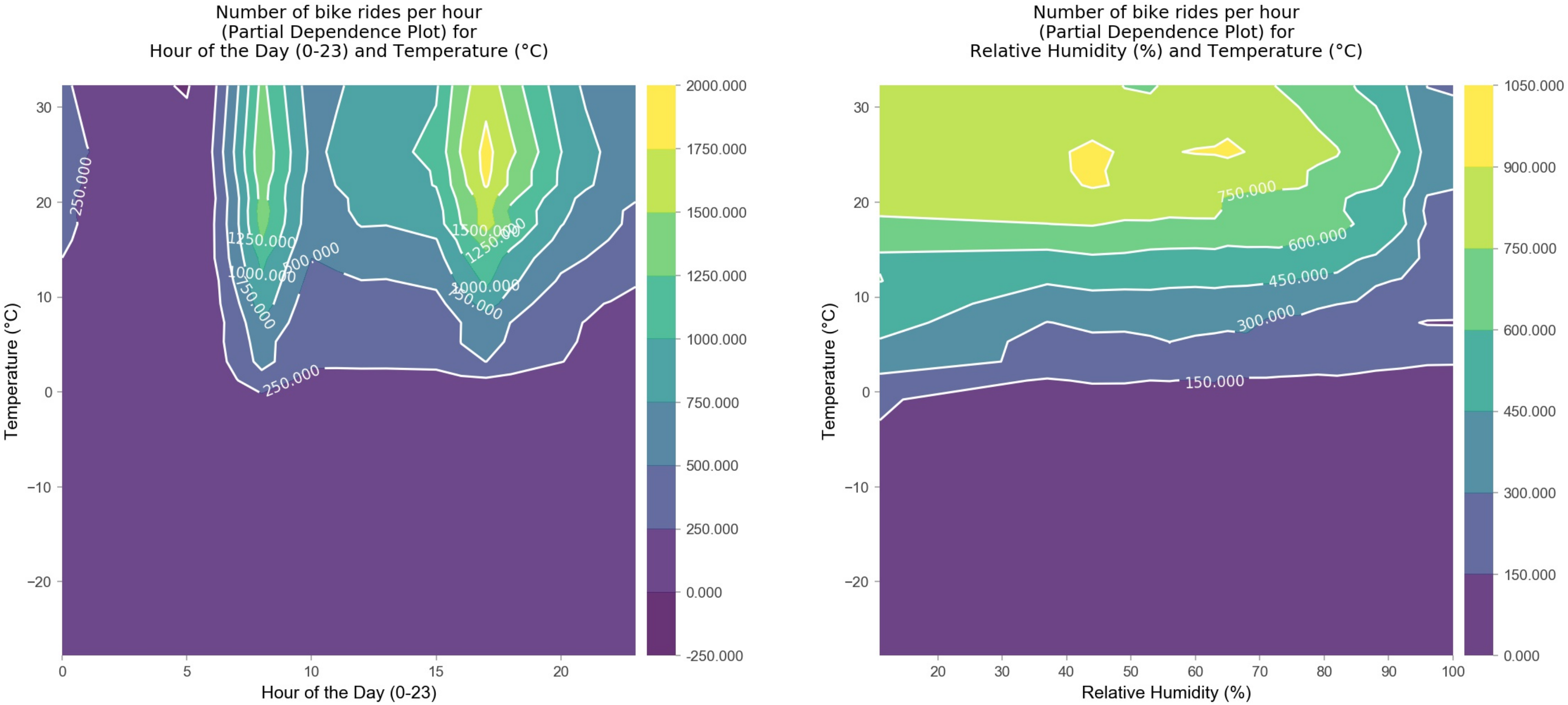


Figure: Partial dependence plots (PDPs) [1,2] for two-way interactions. These plots show the the effect of two variables on the predicted outcome (number of predicted hourly bike rides, coded as the color in the plot). The two features, in this case, are the day of the week and hour of the day. These plots do not show further, higher-order interactions with other features.

Note: The strength of two-way interactions was determined by estimating a gradient boosting regression model similar to the model described before, but with using main effects and two-way interactions as distinct features and then calculating the variable importance.

Findings: Interactions



Number of bike rides per hour
(Partial Dependence Plot) for
Hour of the Day (0-23) and Temperature (°C)

Number of bike rides per hour
(Partial Dependence Plot) for
Relative Humidity (%) and Temperature (°C)

Figure: Partial dependence plots (PDPs) [1,2] for two features. The number of predicted hourly bike rides is coded as the color in the plot. For a more detailed explanation, see first occurence of this plot type on the previous page.

The interaction between **temperature and hour of the day** is the strongest two-way interaction. The plot shows that the main effect (peaks at rush hour) is only valid for temperatures above approx. $0^\circ C$. This also applies for the interaction between **temperature and relative humidity**: The number of predicted bike rides drops with increasing relative humidity only for temperatures above $0^\circ C$. For lower temperatures, there are hardly any rides at all.

Findings: Interactions

Atmospheric pressure has almost no visible main effect (see figure on the bottom), but it turned out to be the second-most important interaction (with hour of the day). This interaction is depicted in the figure on the right: In the morning, number of bike rides increase when atmospheric pressure is above about $100\,kPa$, but in the afternoon, the increase in bike rides seems to happen at a slightly lower atmospheric pressure.

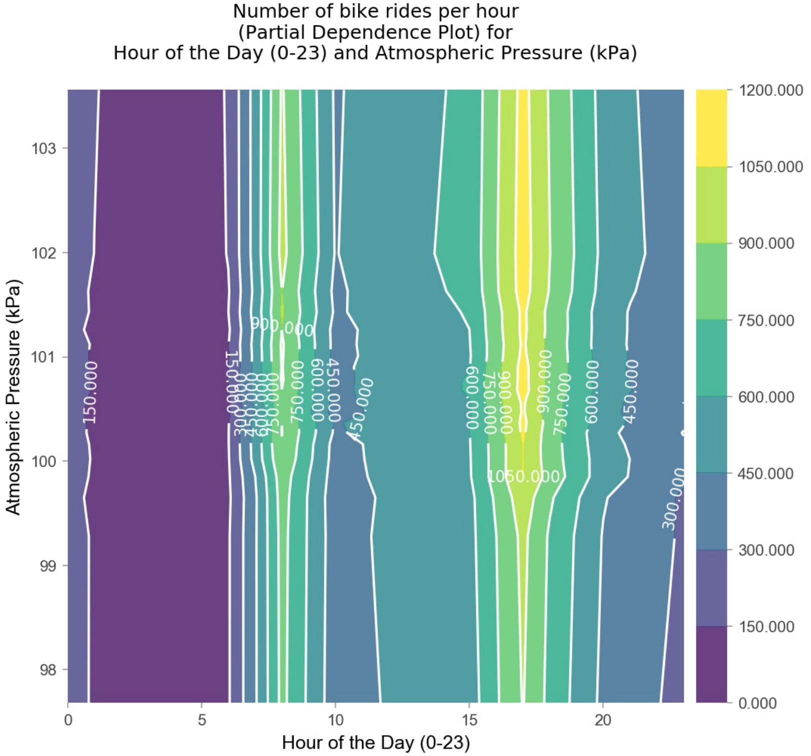


Figure: Partial dependence plots (PDPs) [1,2] for atmospheric pressure and hourh of the day. The number of predicted hourly bike rides is coded as the color in the plot. For a more detailed explanation, see first occurence of this plot type on one of the previous pages.

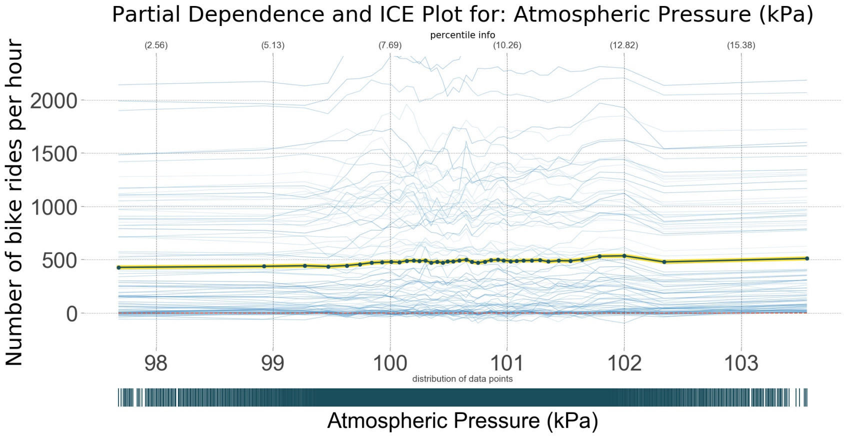


Figure: Partial dependence plot (PDP) shows no obvious main effect of atmospheric pressure on numer of bike rides.

Limitations

There are a few methodological issues that might merrit discussion. First, the gradient boosting model might have slightly overfitted the training data, but still the results in the test that were quite okay. Also, the decision to treat day of the week and time of day as continuous variables (and not categorical) might be worth revisiting (even given that the predictive power was better this way).

Regarding the data used, it might have been beneficial to use precipitation (rain, snow) as additional predictor. Unfortunately, this feature was not available for the given weather station. Moreover, it might also be worth using not actual weather, but weather predictions as additional features. However, this was beyond the scope of this project.

The results identified in this analysis might also be applicable for other cities, but only as long as the climate is somewhat comparable. Probably the biggest limitation of this analysis is that it only considers influences within the bike sharing system (i.e., how many of the available bikes are rented). It does not consider outside influences like traffic, availability of other means of transport, or the need for additional bike renting stations. This sets rather narrow boundaries on the usefulness of these results, unfortunately.

Conclusions

The number of bike rides seems to depend heavily on weather and time variables. It is possible to explain most of the variation and to make somewhat accurate predictions with a machine learning model.

The most important factors are relatively obvious ones. The number of bike rides increases with temperature (picking up at about 2-$3^\circ C$, up to about $25^\circ C$). Atmospheric pressure does not have a strong effect on its own, but interacts heavily with other factors.. For relative humidity, the number of bike rides drops with increasing humidity. Hour of the day might be the most obvious factor, with most bike rides in the morning and in the afternoon.

For all of the variables, there are strong interactions. For example, bike rides only have morning and afternoon peaks during work days, not at weekends, and high relative humidity leads to fewer bike rides in general, but more so for higher temperatures.

In summary, the analysis confirmed obvious assumptions. Probably the biggest surprise was how well the model was able to predict the number of bike rides.

References

1. Molnar C. Partial Dependence Plot (PDP) [Internet]. 2018. Available: https://christophm.github.io/interpretable-ml-book/pdp.html

2. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation [Internet]. 2014. Available: https://arxiv.org/pdf/1309.6392.pdf

3. Wheatley J. Random forest or gradient boosting? [Internet]. 2014. Available: http://joewheatley.net/random-forest-or-gradient-boosting/

4. Parr T, Turgutlu K, Csiszar C, Howard J. Beware Default Random Forest Importances [Internet]. 2018. Available: http://explained.ai/rf-importance/index.html

5. Morgan L. Why Does it Rain When the Pressure Is Low? [Internet]. 2017. Available: https://sciencing.com/rain-pressure-low-8738476.html

6. Haby J. RELATIVE HUMIDITY PITFALLS [Internet]. Available: http://www.theweatherprediction.com/habyhints2/564/

7. Dotson JD. How Temperature & Humidity are Related [Internet]. 2018. Available: https://sciencing.com/temperature-ampamp-humidity-related-7245642.html

8. Drury M. Cross Validated: Relative variable importance for Boosting [Internet]. 2017. Available: https://stats.stackexchange.com/questions/162162/relative-variable-importance-for-boosting