**To:**

Dr. James E. Hetrick

Professor and Fletcher Jones Endowed Professor of Data Science

**From:**

Ihor Markevych

University of the Pacific Graduate Student

**Report on “Bikeshare” project**

**Task:**

The goal of conducted analysis is to create two models that would be able to predict bike share usage in for Capital Bikeshare system, Washington D.C., USA. First model should be able to work with daily data, while second should work on hourly bases data.

**Data:**

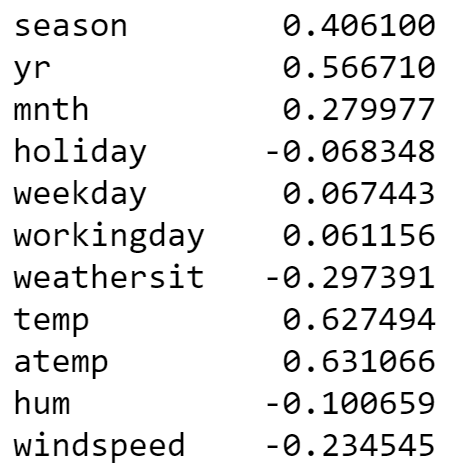
Data that was used for this project is coming from *Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.* Data is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is publicly available in [*http://capitalbikeshare.com/system-data*](http://capitalbikeshare.com/system-data). The data on two hourly and daily basis and then extracted and added the corresponding weather and seasonal information. Weather information was extracted from [*http://www.freemeteo.com*](http://www.freemeteo.com). Data has 17 features for hourly based data and 16 features for daily based data. Data contains next features:

* + record index;
  + date;
  + season (springer, summer, fall, winter);
  + year (2011 or 2012);
  + month (1 to 12);
  + hour (0 to 23);
  + weather day is holiday or not;
  + day of the week;
  + weather day is workingday or not;
  + weather;
  + normalized temperature in Celsius;
  + normalized feeling temperature in Celsius;
  + normalized humidity;
  + normalized wind speed;
  + count of casual users;
  + count of registered users;
  + count of total rental bikes including both casual and registered.

**Initial analysis:**

Upon exploratory data analysis strong correlations between several indicator and goal variable were found.

**For daily data:** **For hourly data:**

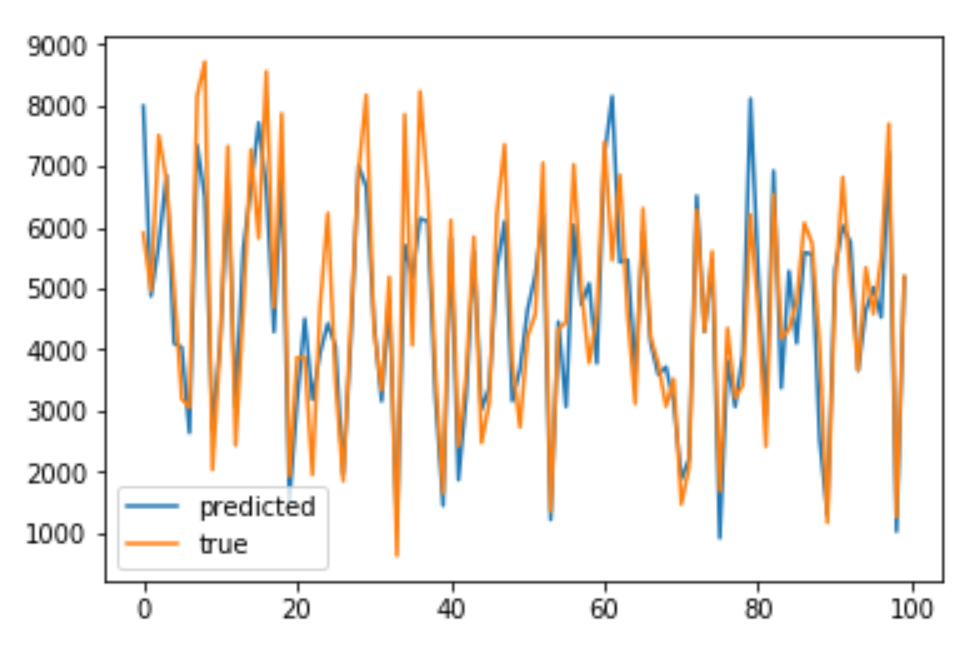
 

**Linear regression models:**

Linear regression models were built for both hourly and daily data. Both models were evaluated on unseen during training data.

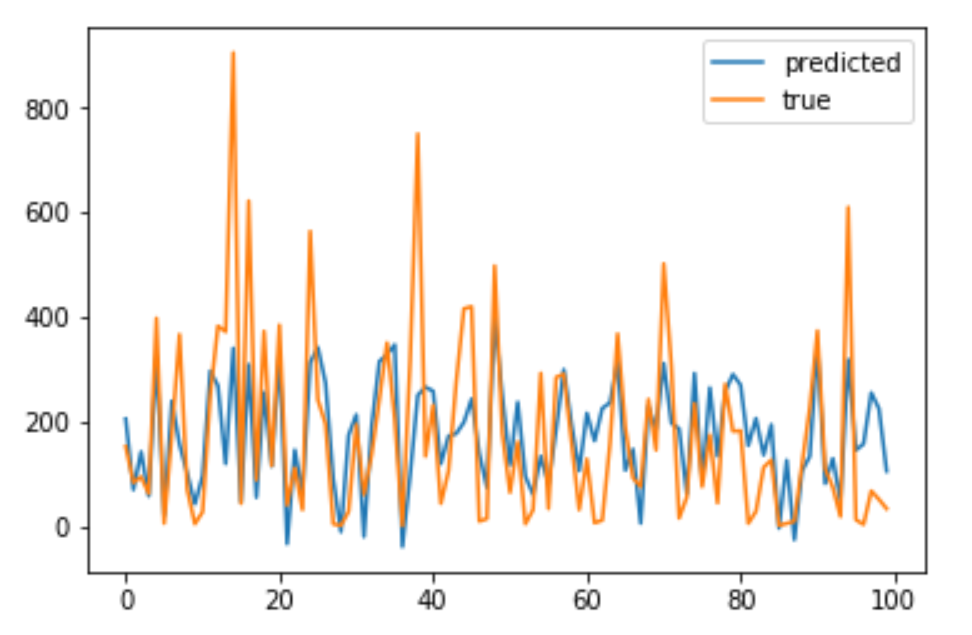
***Daily data:***

Daily data appeared to be more predictable with linear regression, with 80% of variance explained by the model and with mean absolute error of 669 bike shares. As it can be seen per plot below, model’s predictions closely follow true data. Next variables gave best performance: *season, year, month, weather, temperature, feeling temperature, wind speed, humidity*.



***Hourly data:***

Linear regression on hourly data gave significantly worse results, with only 40% of variance explained by the model and with mean absolute error of 105 bike shares. As it can be seen per plot below, model fails to predict peak usage. Best performance was achieved if all predictors were included in the model.

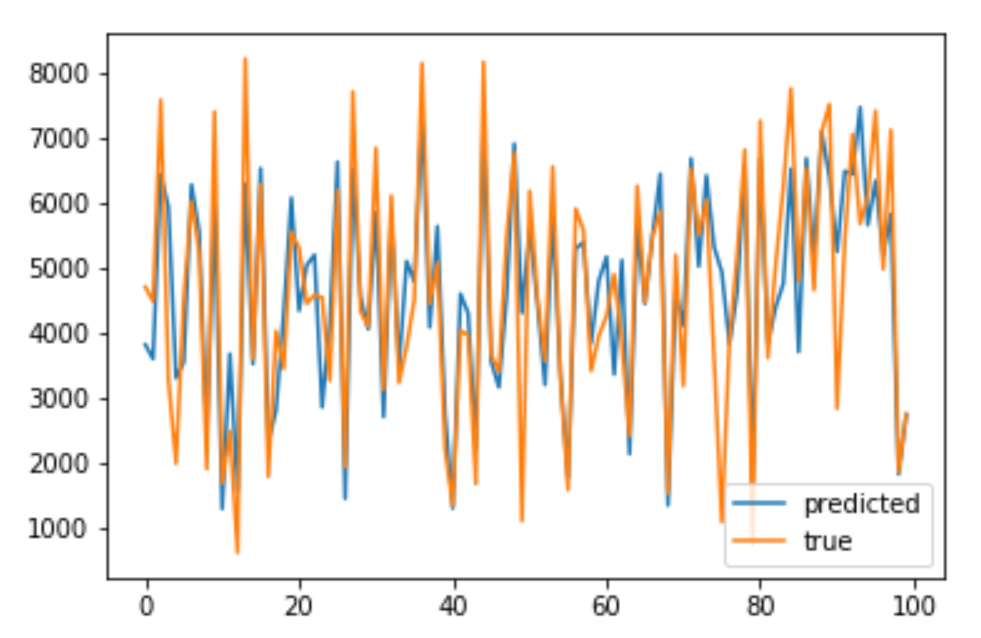


**Dense Neural Networks:**

Additionally two dense neural network models were built on each of the datasets available. As in case with linear regression, both models were evaluated on unseen during training data.

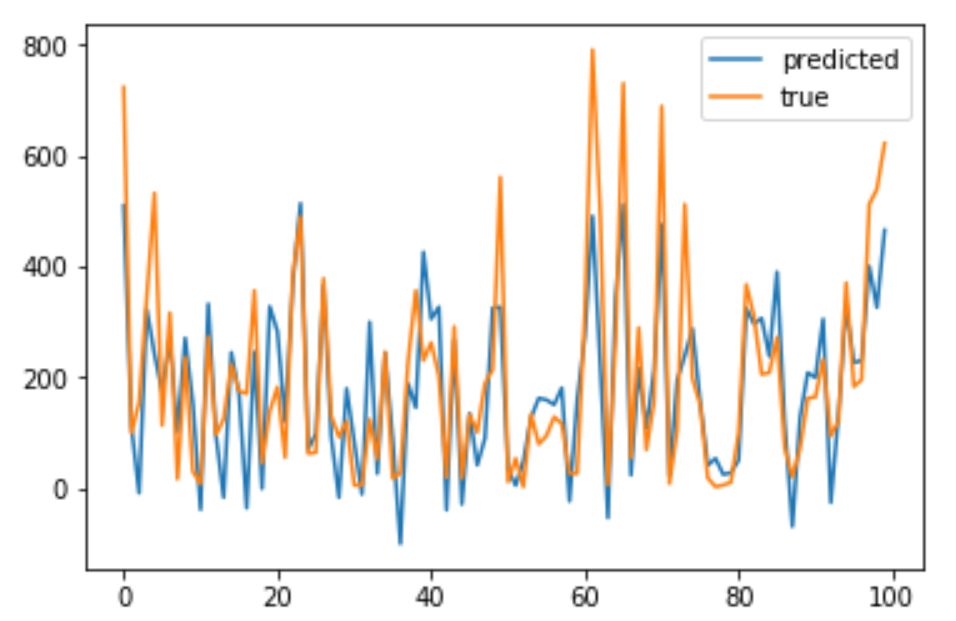
***Daily data:***

Dense neural network gave comparable to linear regression performance, with 75% of variance explained by the model and with mean absolute error of 726 bike shares. As it can be seen per plot below, model’s predictions closely follow true data. All variables were included in the model.



***Hourly data:***

Dense neural network gave significantly better results in comparison with linear regression, with 62% of variance explained by the model and with mean absolute error of 81 bike shares. As it can be seen per plot below, model follows data much closer. All variables were included in the model.



**Conclusion:**

It is proposed to use linear regression model for daily data, as it is computationally less expensive and, more important, it is explainable, as its coefficients allow to explore found relations between predictors and goal variable.

For hourly data either of the model can be used, depending on whether accuracy is more important (neural network), or explainability is crucial (linear regression).