Hourly Usage Prediction for Capital Bikeshare in Washington,D.C

Hangyu Ren, Joel McGuire, Kegan Schifferdecker

Bike Sharing Programs

Bike sharing programs are services where bicycles are made available for shared use to people on a short time period basis for a price.

Today, there are over 800 bike-sharing programs around the world (Fishman, 2016)

Fishman, Washington, & Haworth, (2014) found that bikeshares in Washington D.C. significantly replaced automobile use, reducing congestion and pollution in the city.

A critical problem for bike-share systems is finding the optimal way to redistribute bikes from destination nodes to embarcation nodes (Chemla et al, 2013, Schuijbroek et al, 2017). While predictive analytics does not directly solve this problem, its integration with optimization methods is deemed promising.

Goal of the project

Based on the data we have, we are going to use predictive analytics method to forecast how many bikes are needed per hour, per day even per week.

If the bike usage can be predicted on the granularity of an hour, it can assist in the optimization of repair and redistribution schedules.

Data Fields

(Variables or Features)
Data Source :
https://www.kaggle.com/c/bike-sharing-demand

datetime - hourly date + timestamp

season - 1 = spring, 2 = summer, 3 = fall, 4 = winter

holiday - whether the day is considered a holiday

workingday - whether the day is neither a weekend or holiday

weather - [**1**: Clear, Few clouds, Partly cloudy, Partly cloudy , **2**: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist **3**: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds , **4**: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog]

temp - temperature in Celsius

atemp - "feels like" temperature in Celsius

humidity - relative humidity

windspeed - wind speed

casual - number of non-registered user rentals initiated

registered - number of registered user rentals initiated

count - number of total rentals (This is what we are trying to predict, the observations are hourly).

Data Preparation

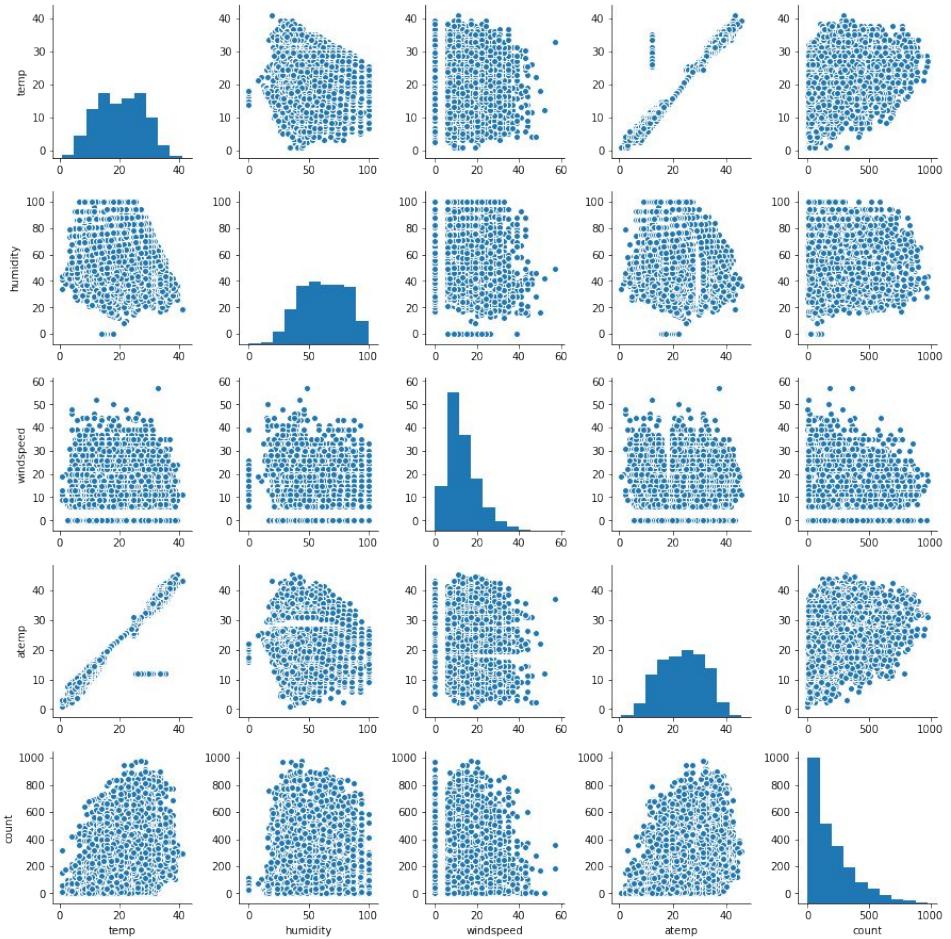
- To prepare the data for modeling, we needed to analyze the data and make changes that would cause problems with the models creation and results
  - Checked if there were any missing values in the data
  - Change the datetime column given in data into individual month, week, day, hour columns
  - Change datatype of weather, holiday, working day, and season columns into categorical data types since data is given in categories

Exploratory Analysis

- After the data has been prepared for modeling, the next step is to do exploratory analysis on the data to find issues where the variables in the data could cause an increased error in the predictive accuracy of the models.
  - See how the different variables (data columns) interact by using pair plot
  - Determine correlation values between the columns

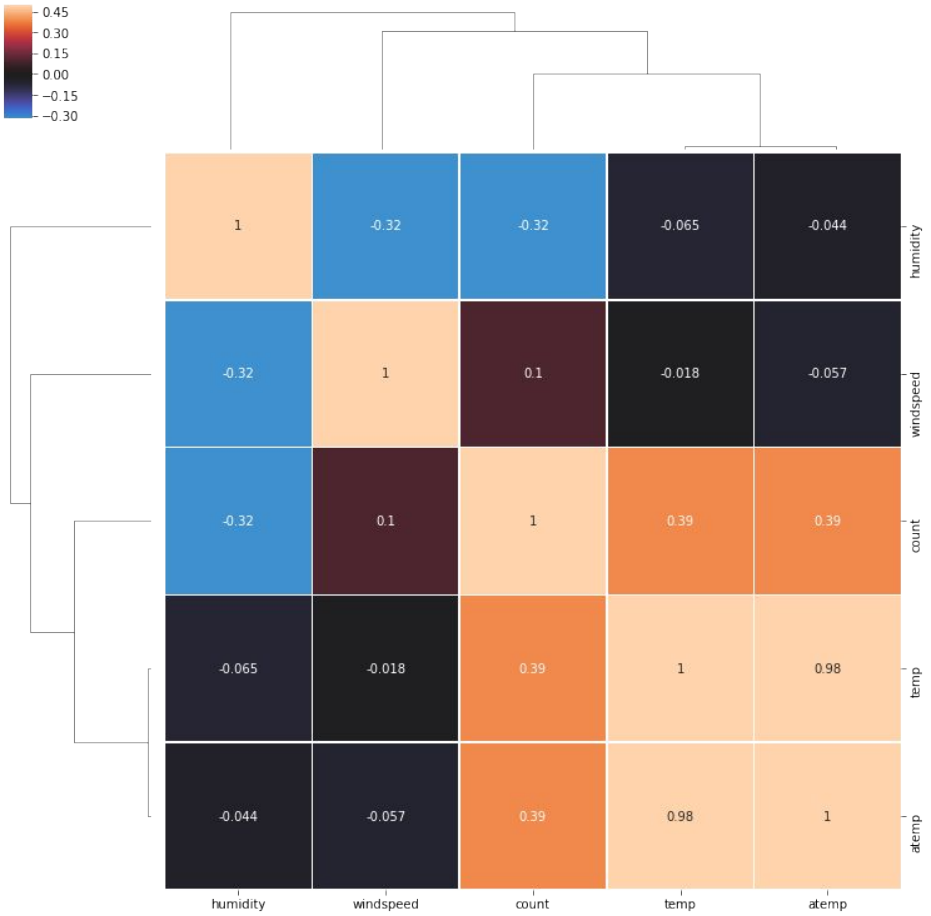
sns.pairplot(train[["temp", "humidity", "windspeed", "atemp", "count"]])

We see that wind speed and count are roughly exponentially distributed, it may be helpful to do a log transformation of these variables.



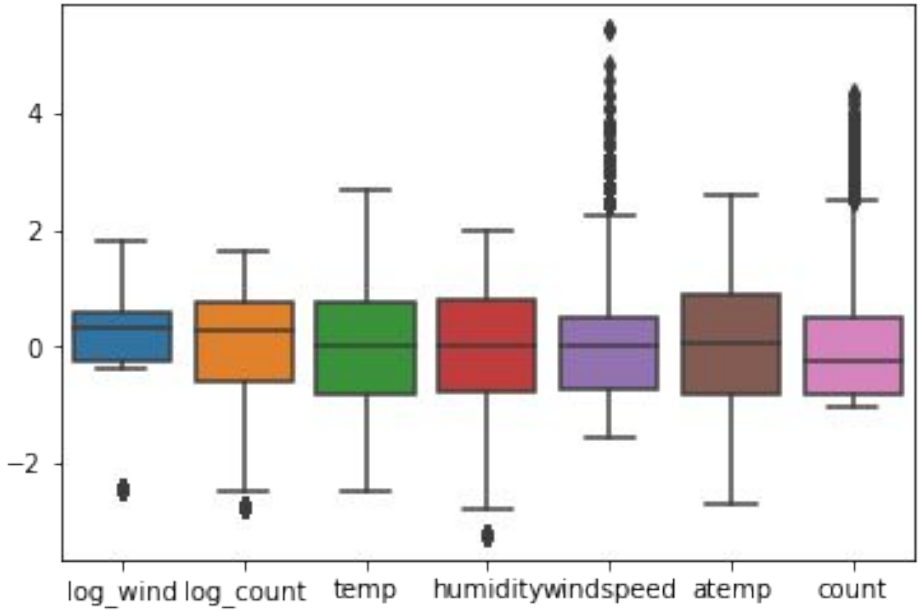
```
sns.clustermap (train[["temp", "humidity", "windspeed",
"atemp", "count"]].corr(), center=0, linewidths=.5,
figsize=(13, 13),vmax=.5, square=True,annot=True)
```

Superficially, it appears as if temperature and temperature “feel” has the strongest relationship to ride count. It’s positive, which makes sense.



Outliers

We have quite a few outliers for *windspeed* and *count* which appear to be exponentially distributed, but applying a log transformation to them appears to make them behave more nicely.



Code

```
numeric_vars = train[["temp", "humidity", "windspeed", "atemp",
"count"]]
log_num = numeric_vars[["windspeed", "count"]].apply(np.log1p)
log_num.columns = ['log_wind', 'log_count']
log_num
numeric_vars = pd.concat([log_num, numeric_vars], axis = 1)
numeric_vars
numeric_vars = preprocessing.scale(numeric_vars)
numeric_vars = pd.DataFrame(numeric_vars)

numeric_vars.head()
numeric_vars.columns = ["log_wind","log_count", "temp", "humidity",
"windspeed", "atemp" , "count"]

sns.boxplot(data = numeric_vars )
```

Modeling Strategy

Feature Engineering: If riding a bike instead of a car is an aspirational activity, research suggests that extracting temporal features that capture the start of new periods (week, month, year) will add accuracy to ridership prediction (Dai et al, 2014).

Data Splitting: Since the test data is selected non randomly (it's the last days of the month) there's reason to believe that the test and training data aren't i.i.d. This can be tested by seeing if our cross validated prediction of which dataset observations datasets belong to is better than random.

Models: We employ a modest ensemble of linear models, k-nearest neighbors, and tree methods, with a random forest super learner to predict hourly bike share demand.

Current Results

The competition is judged on who makes the best out of sample prediction measured as lowest root mean squared logarithmic error.

Our model : 0.5596.

Benchmark: 1.58455

There is room for improvement, but our model is a definite improvement over the benchmark.

Improvements that Can Be Made

- Utilize an ensemble if given enough time for computation.

Our code can be viewed and forked from github.
https://github.com/JPMOS/Bikeshare

Please feel free to replicate our results or make any comments.

Conclusion

We are currently 1,807 out of 3,251 teams, suggesting our model has some merit.

References

Chemla, D., Meunier, F., & Calvo, R. W. (2013). Bike sharing systems: Solving the static rebalancing problem. *Discrete Optimization*, *10*(2), 120-146.

Schuijbroek, J., Hampshire, R. C., & Van Hoeve, W. J. (2017). Inventory rebalancing and vehicle routing in bike sharing systems. *European Journal of Operational Research*, *257*(3), 992-1004.

Fishman, E., Washington, S., & Haworth, N. (2014). Bike share's impact on car use: Evidence from the United States, Great Britain, and Australia. *Transportation Research Part D: Transport and Environment*, *31*, 13-20.

Dai, H., Milkman, K. L., & Riis, J. (2014). The fresh start effect: Temporal landmarks motivate aspirational behavior. *Management Science*, *60*(10), 2563-2582.

Kaggle.com. Bike Sharing Demand. https://www.kaggle.com/c/bike-sharing-demand, 2018.

Fishman, E. (2016). Bikeshare: A review of recent literature. *Transport Reviews*, *36*(1), 92-113.