draft1

Ricky Heinrich & Vimaljeet Singh

2023-03-24

Introduction

The dataset we have chosen to analyze is Hadi Fanaee-T's Bike Sharing Dataset, from the Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto, accessed in the UCI Machine Learning Repository. This dataset combines the Trip History Data for the years of 2011 and 2012 of 'Capital Bikeshare', which is metro Washington DC's bikeshare service, with weather data and the holiday schedule. We are hypothesizing that we are able to predict the count of bikes rented in a given hour given the predictor variables (weather, calendar date and time, etc). We are also hypothesizing that results will be better if we model casual user bike rentals separately than registered users. Finally, we want to investigate which predictor variables are most important in making the predictions.

Description of Dataset

The data consists of an aggregated count of 'rides' by hour, over the span of the years 2011 and 2012. It contains 17379 rows and 17 columns.

Variable Name	Description	Type
instant	Record index	ordinal
dteday	Date	datetime
season	Season (winter, spring, summer, fall)	categorical
yr	Year (2011, 2012)	ordinal
mth	Month	categorical
hr	Hour	categorical
holiday	Whether day is a holiday or not	boolean
weekday	Day of the week	categorical
workingday	If day is neither weekend nor holiday	boolean
weathersit	Weather conditions	ordinal
temp	Temperature in Celsius	numerical
atemp	Feeling temperature in Celsius	numerical
hum	Humidity	numerical
windspeed	Wind speed	numerical
casual	Count of bikes rented by casual users	numerical
registered	Count of bikes rented registered users	numerical
cnt	Count of total bikes rented	numerical

Due to high colinearity, we've decided to remove the `atemp` variable as well as the `workingday` variable from our modelling. We are considering our full model to include the following predictors: yr, mth, hr, holiday, weekday, weathersit, temp, hum, and windspeed. The response variables are cnt, casual, and registered.

Plan overview

We will be conducting a linear regression, as it is the most simple model, as well as investigating a poisson regression, since this is technically count data, and a random forest model, to see if it improves the predictions. We'll investigate these models for the total count, as then separately for the casual counts and registered counts. We will investigate outliers in the data, and we will use variable selection methods to investigate which variables are most important. We are separating our data into an 60/40 training/testing set, and we will be using MSE as a measure of fit, as well as Rˆ2. Finally, we will make our conclusions, and talk about limitations and possible future work.

Random Forest model

Because we are interested in prediction, we are ready to lose some interpretability in our model in exchange for better predictive power. Using random forest over a simple decision tree decreases the variance as well as the bias, usually resulting in closer predictions.

Total Count

As seen in Figure ##, the percent increase in MSE shows that when either hr or yr are permuted, then the MSE increases by over 150%: the values of these variables are really important in prediction. Weekday and temperature values are also important, as MSE can increase by 75 or 100% as well when they get jumbled up. All variables except for the windspeed seem to sit at %IncMSE values of over 25%, suggesting that they all carry some importance in the model. [I don't understand the plot vs the raw values, why raw temps has greater %IncMSE than yr but in plot yr comes up first?]

"mean decrease of accuracy in predictions on the out of bag samples when a given variable is permuted" "measure of the total decrease in node impurity that results from splits over that variable, averaged over all trees"

The IncNodePurity shows that the hr variable is by far the most important variable in regards to decreasing the node impurity. [What's the scale on these values idk . . .]

Casual Count

Modeling for just the Casual users, we see in Figure ## that the most important variable in terms of the mean decrease of accuracy in predictions on the out of bag samples when the variable is permuted is now weekday, at over 200%, followed by hr just below 150%, and then a year just under 100%. Still, there's a slew of variables that change over 50% in MSE when they are permutated.

In terms of IncNodePurity, the hr variable still ranks the highest, but the value is not as high as in the total cnt, and the weekday and rawtemp values follow closer behind than other models.

Registered Count

Modeling for just the registered users follows more closely to the total count, which makes sense given that registered users take up a greater proportion of total rides than casual users. We see that hr, year, and weekday are again top variables when it comes to %IncMSE, and that hr is vastly more important when it comes to IncNodePurity.

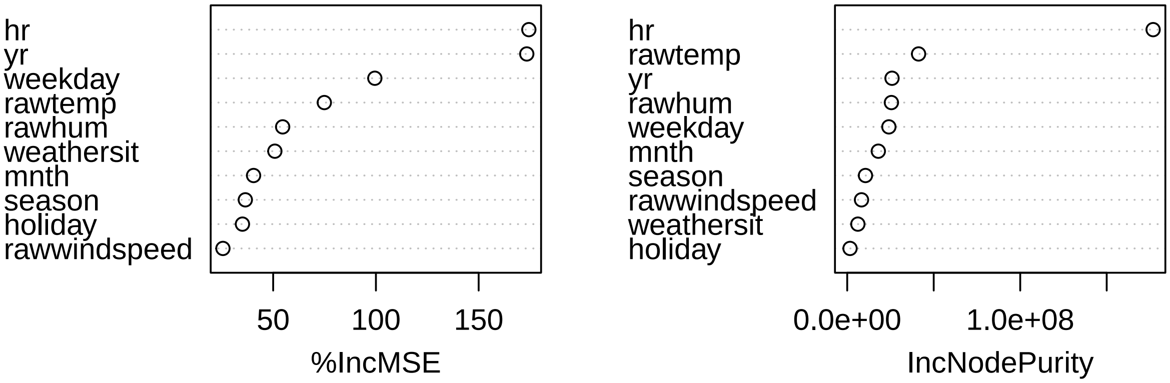

Figure 1: Variable Importance for Total Count

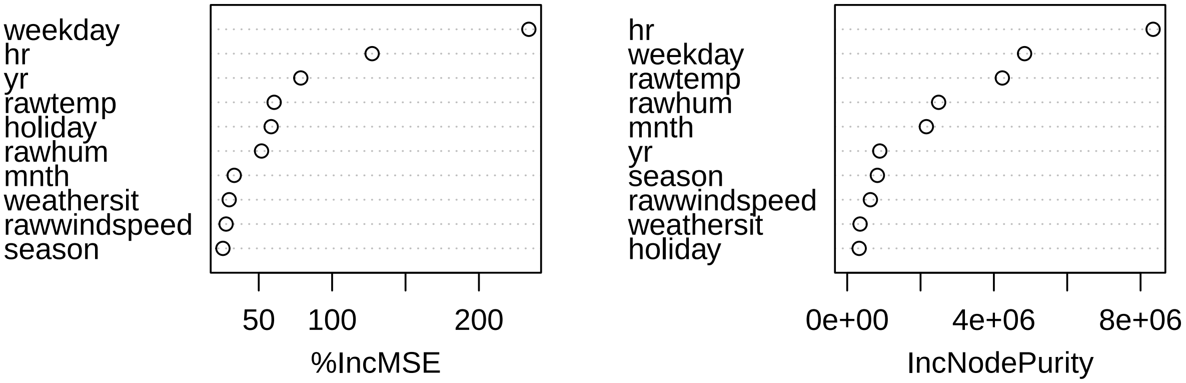

Figure 2: Variable Importance for Casual Count

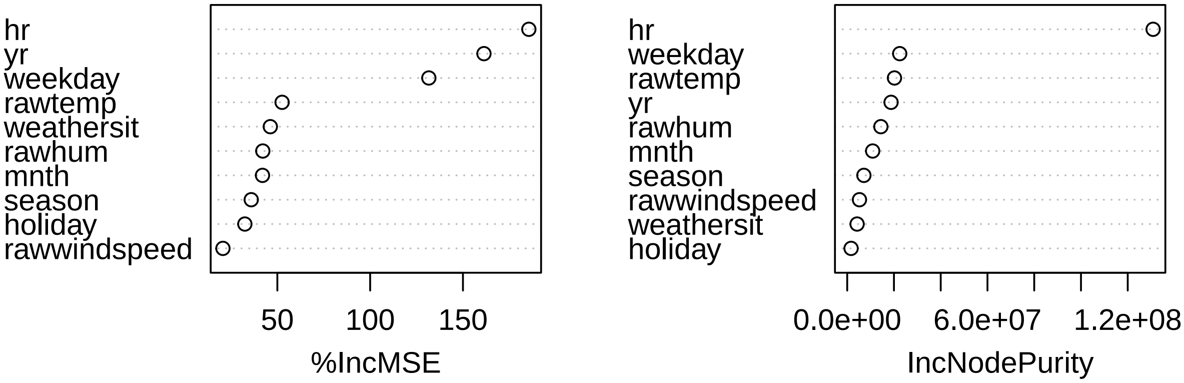

Figure 3: Variable Importance for Registered Count

Perfomance of all models on test set

The final draw: comparing the model performances on the test set. We've calcualted the MSE for each model, and here are the results:

[predtable] MSE performance for all models, maybe have test and train?

We see that 'yr' is an important variable all around: it makes sense, as Capital Bikeshare began operations in 2010, and so usage growth would grow a lot in the first few years. To predict count of bike rentals per hour in future years, we would have to take into account the growth in the business and popularity. Therefore, a limitation in our data is that is only covers two years and so it is hard to forecast usage in future years.

Limitations