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 R^2. Finally, we will make our conclusions, and talk about limitations and possible future work.