UNIVERSITY of WASHINGTON

Forecasting Rental Demands of Bike-Sharing Systems

Group 1

Wei Guo, Yi He, Yan Jin, Xiao Tan

Outline

1. Background
2. Problem Description and Proposed Solutions
3. Data Analysis
4. Results
5. Discussion

Background

- Bike-sharing is becoming popular
  - Sustainable and environmentally friendly
  - Part of public transportation, more and more important
  - Over 500 bike-sharing programs worldwide

- How it works
  - Distributed network of stations
  - Rent and return at any open station
  - Convenient for both registered and casual users
  - Large assortment of related data automatically collected

Problem Description

Objective: Forecast accurate bike rental demands for the given dates based on historical patterns and weather data

Field	Data Type	Description
dteday	date	date from 01/01/2011 to 12/31/2012
season	categorical	1 = spring, 2 = summer, 3 = fall, 4 = winter
yr	categorical	0 = year 2011, 1 = year 2012
mnth	categorical	month, 1-12
weekday	categorical	day of the week; 1-6 = Monday-Saturday, 0 = Sunday
hr	categorical	hour, 0-23
holiday	categorical	0 = not a holiday, 1 = holiday
workingday	categorical	0 = not a working day, 1 = working day
weathersit	categorical	1 = clear, few clouds, 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	continuous	normalized temperature in Celsius; the values are divided by 41 (max)
atemp	continuous	normalized "feels like" temperature in Celsius; the values are divided by 50 (max)
hum	integer	normalized humidity; the values are divided by 100 (max)
windspeed	continuous	normalized wind speed; the values are divided by 67 (max)
casual	integer	number of non-registered user rentals
registered	integer	number of registered user rentals
cnt	integer	number of total rentals

Proposed Solutions: Count Models

Poisson Regression	**Negative Binomial Model**
$\log(\mu) = \beta_0 + \beta_1 X_1$	$\log(\lambda) = \beta_0 + \beta_1 X_1 + \varepsilon$
Assumptions 1. Independently distributed 2. Mean and variance is equal to μ	Assumptions 1. Independently distributed 2. $\exp(\varepsilon_i)$ is a gamma-distributed error

Principle Components Analysis

- Reduce number of variables
- Reduce multicolinearity
- Simplify redundant information
- Reduce the complexity of large sets of correlated variables.

Proposed Solutions: Ensemble Learning

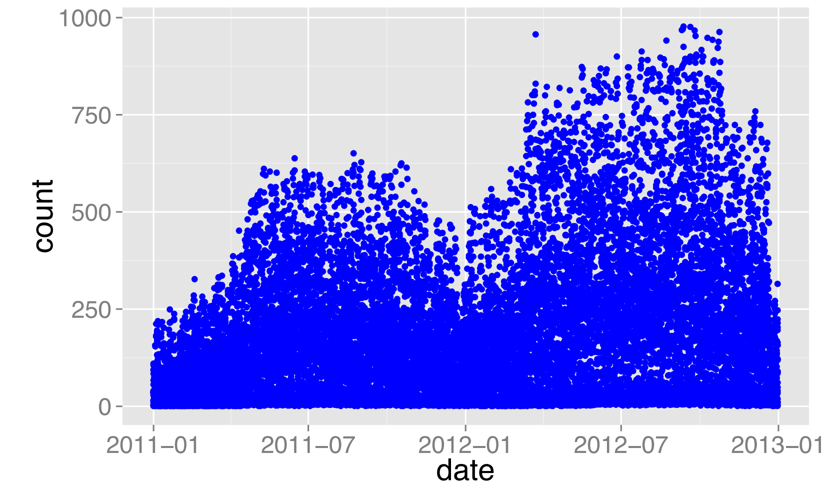
Benefits of Ensemble Learning (decision tree based method)

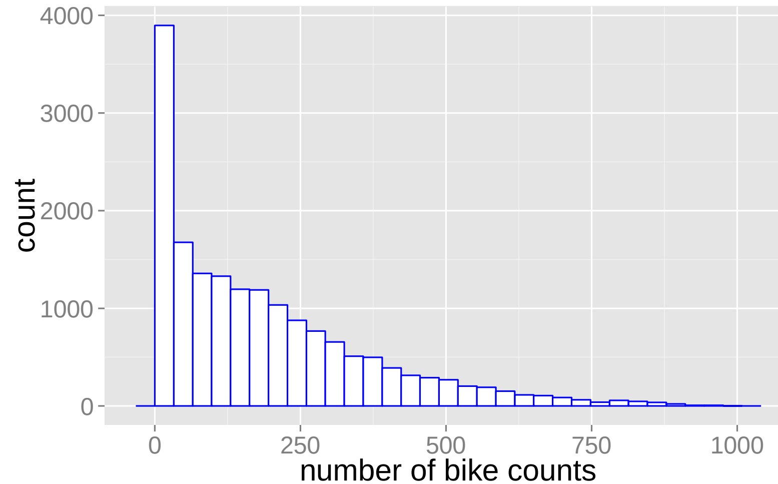
- Makes no assumptions (non parametric)
- Can handle different types of variables
- Robust to over-fitting
- Deals with missing data well
- Relatively easy to use
- Overall good performance

Random Forest	**Gradient Boosting Method**
Bagging Method	Boosting Method
Low variance, high bias	Low bias, high variance
Run in parallel	Run sequentially

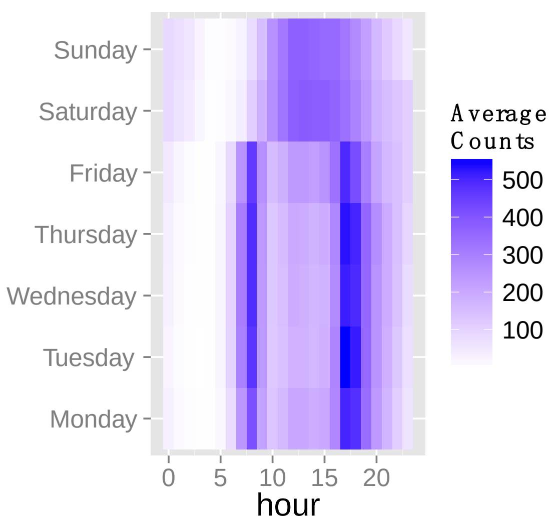
Data Analysis

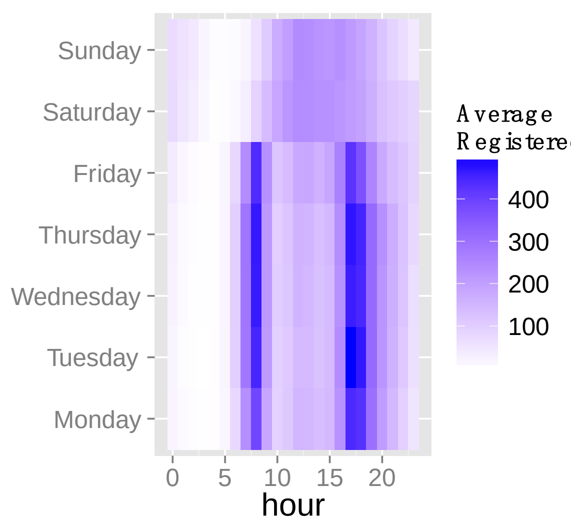
- Distribution of count variable

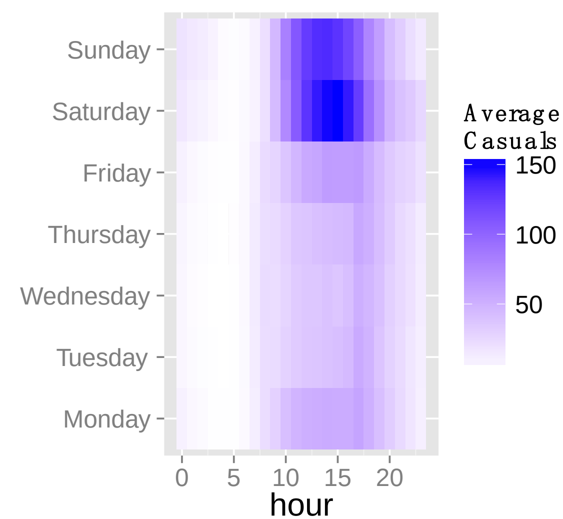




- User behavior analysis

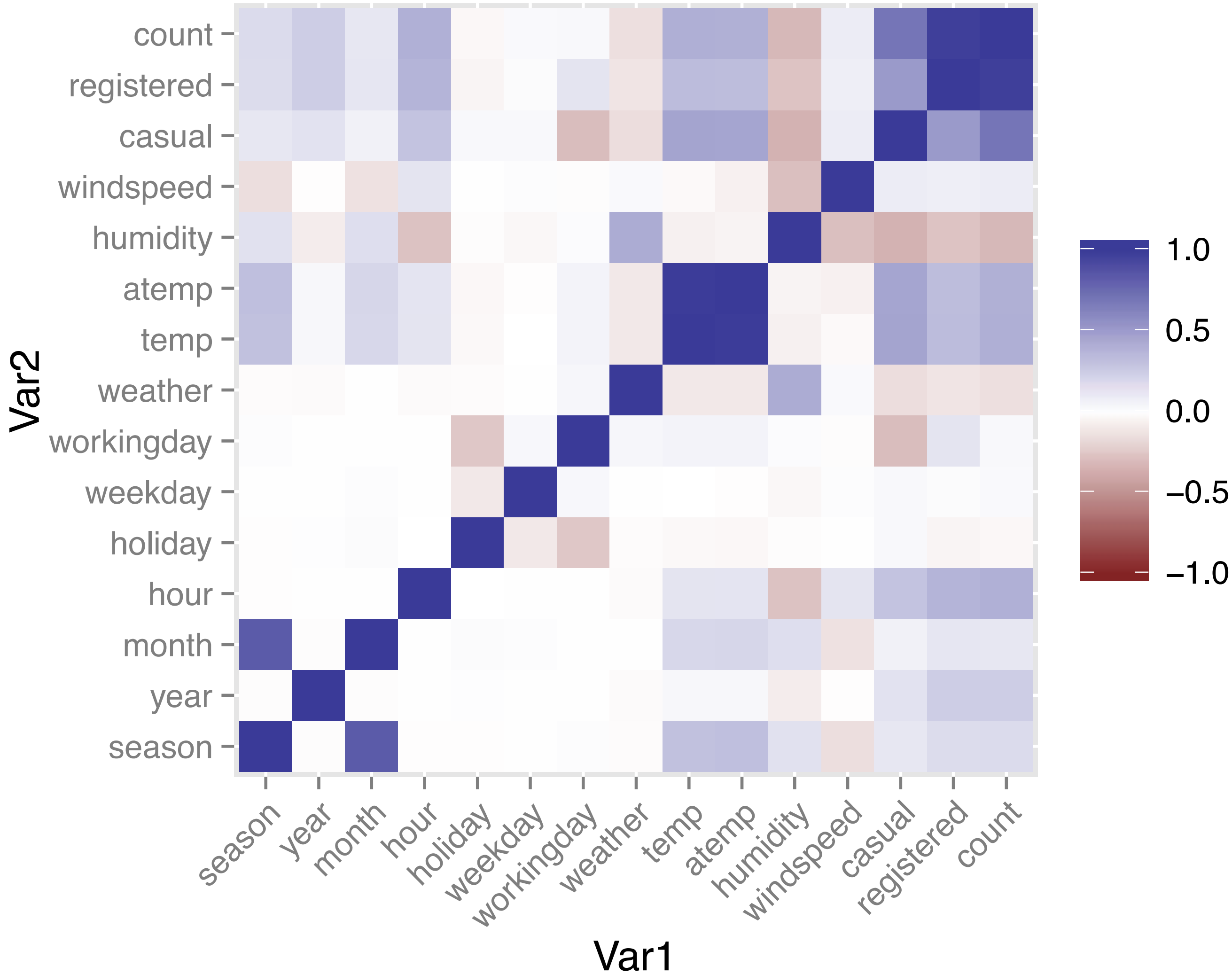




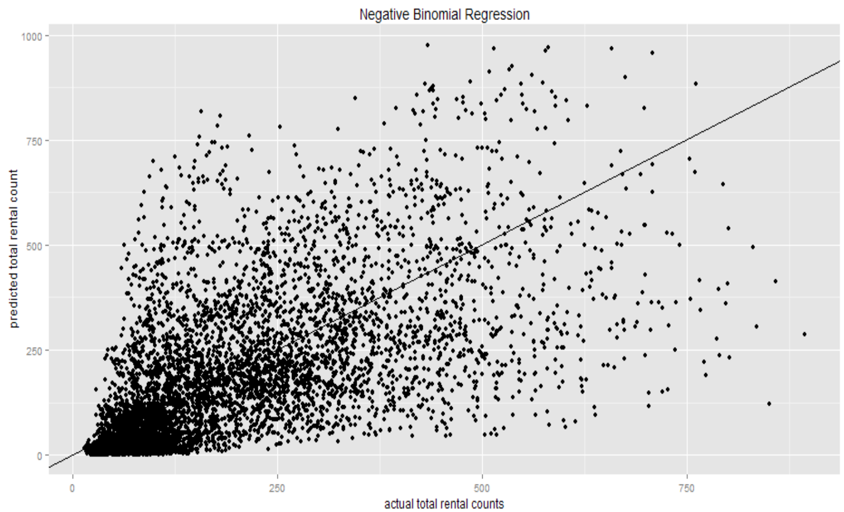


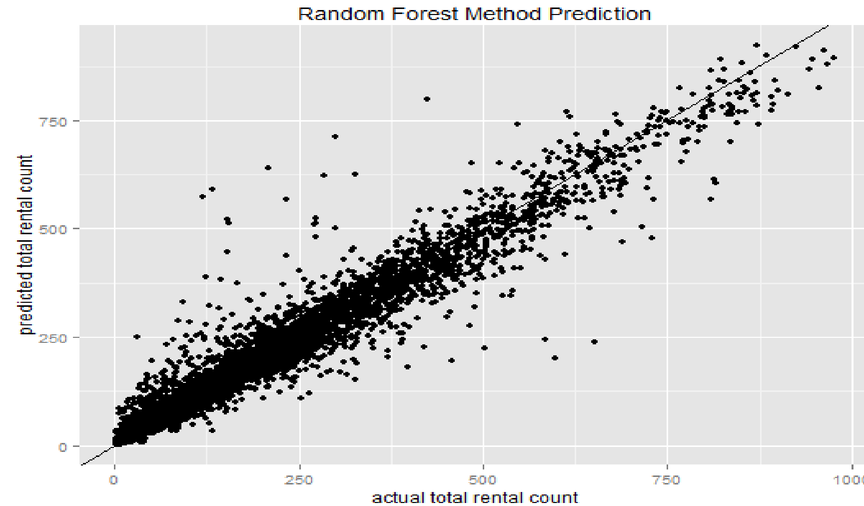
Data Analysis

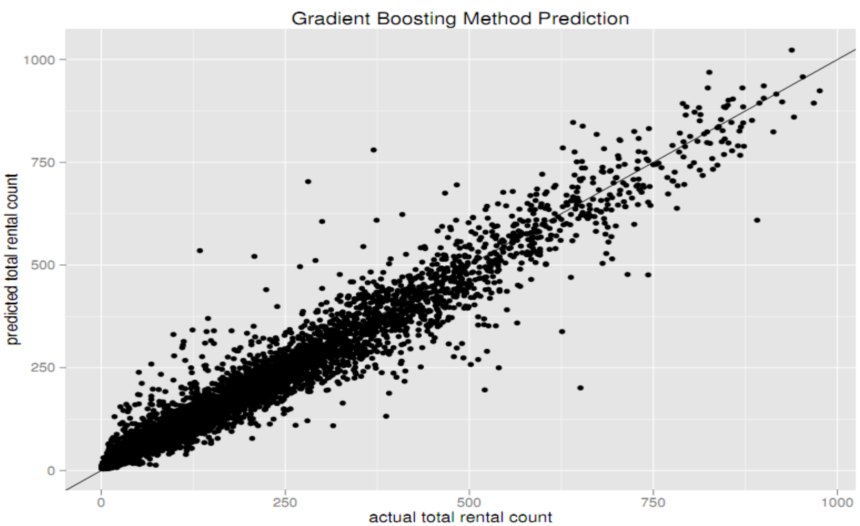
- Correlation



Estimation results

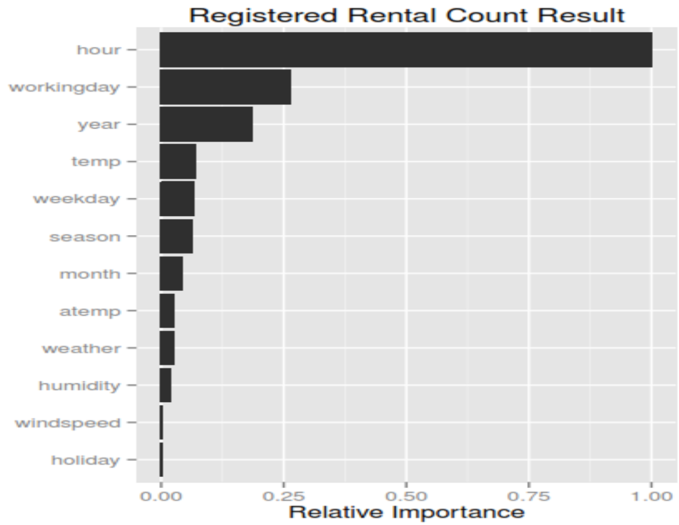


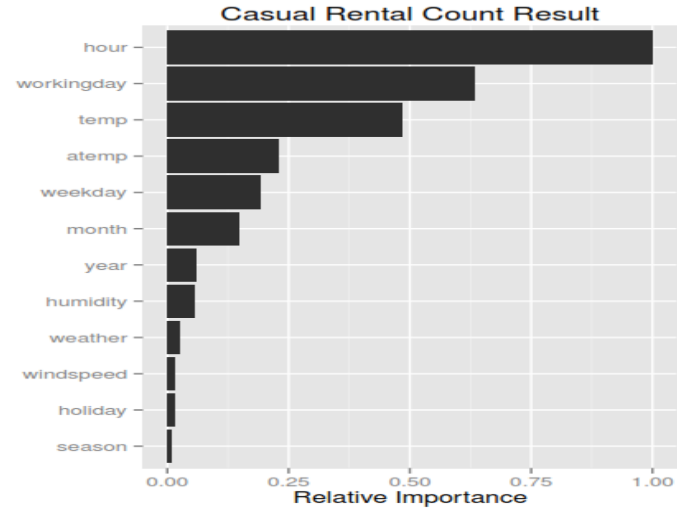


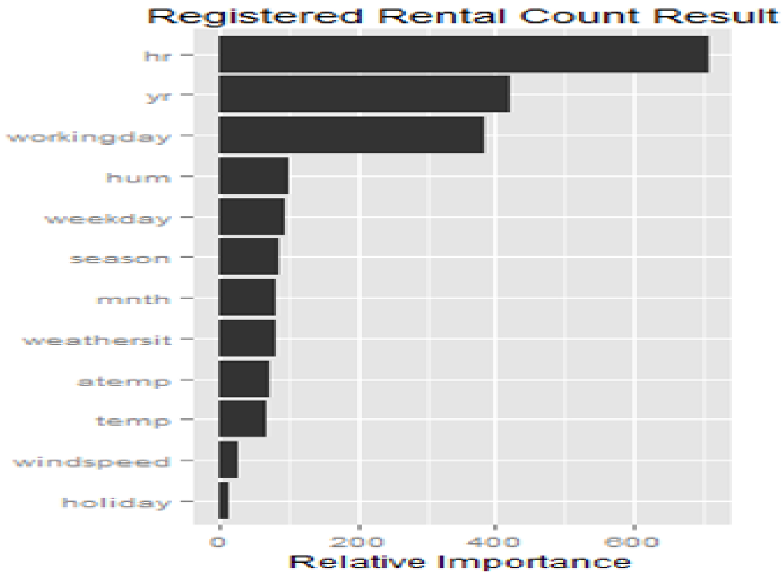


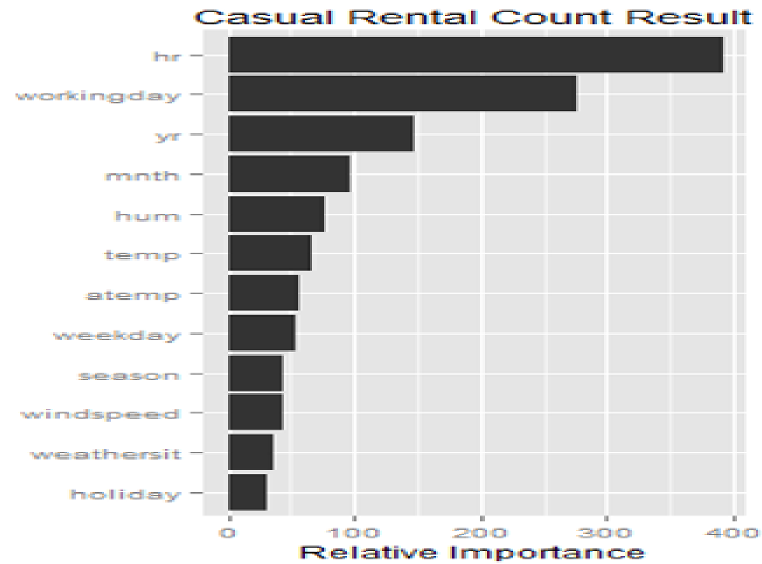
Method	Total RMSLE
Poisson	1.734
Negative Binomial	1.107
Random Forest	0.329
Gradient Boosting	0.403

Feature selection









Discussion

- Random Forest and Gradient Boosting outperform the Poisson based regression
- Several characteristics of user behavior are identified
- Casual and registered count models are separately trained
- Feature selection is a significant source of improvement in predictions
- PCA may not improve the accuracy of models
- Future research: Time series analysis, relative-importance of regressors in Negative-Binomial regression model