Bike Sharing Analysis

Qinmiao Deng* Kira Shen* Xinbei Yu* Kaishuo Zhang* Qi Zhao*

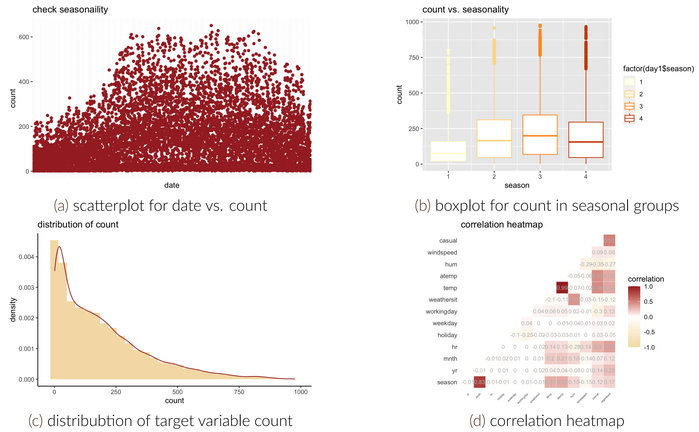
Brown University

Introduction

Bike sharing systems offer a seamless, automated process from membership to rental and return. By analyzing data on user behavior, we can gain insights into usage patterns and user preferences. This project utilizes the UCI Machine Learning Repository's bike sharing dataset, consisting of 17 features and 17379 data points, to predict the total number of rental bikes (cnt in the dataset). Our analysis aims to provide valuable insights for bike sharing providers and urban planners, helping to optimize the system for users and promote sustainable transportation.

Exploratory Data Analysis

The variable instant, which track the position of each data point in the document, is dropped as it records the same information as default index. The variable dteday, which measures date of each data point documented is also dropped as seasonality shown below is not strong. Variable registered and casual is combined into cnt, and then dropped. Based on the correlation heatmap, feature apparent temperature is dropped from the model as it's 0.99 correlated with temperature Target variable cnt shows a right-skewed distribution, and then is log-transformed in later analysis.



(a) scatterplot for date vs. count (b) boxplot for count in seasonal groups

(c) distribubtion of target variable count (d) correlation heatmap

Figure 1. EDA

Methods

1. **Linear Regression**, linear regression has high interpretability of the linear regression model. To be specific, unlike other complex model or dimension reduction method, the simple linear regression model provides information on the important features that contribute to our prediction.
2. **Principal Component Analysis**, PCA is an unsupervised learning approach to reduce the dimensionality of a dataset. PCA transforms the original covariates into abstract components of less dimensions but retain as much information as possible.
3. **Lasso Linear Regression**, Lasso shrinkage helps addressing the potential overfitting of training data, and it also presents a clear feature importance.

Simple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

Target variable cnt is log-transformed because qq plot suggests a positive-skewed distribution, and a skewed distribution will violate the normality assumption of linear regression.

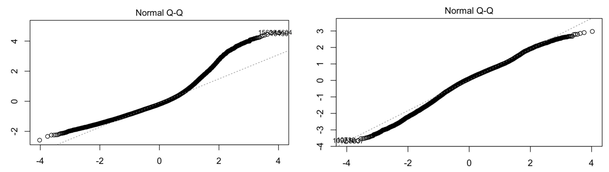


Figure 2. qq plot before and after log-transformation

Forward feature selection is used because the number of features are relatively small. The final model is consist of season, year, hours, holiday, working day, temperature, humanity, windspeed, and whether situation.

This project uses R^2 as evaluation metrics. The second method is principal component analysis(PCA), and proportion of variance explained is a key metrics, making the comparison between the two model easier

The model is also cross validated with dataset being split in 80-20.

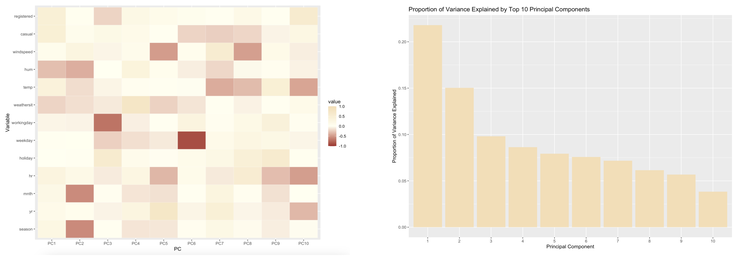
Linear Regression with Dimension Reudction - PCA

PCA is first performed on the input matrix and then linear regression is performed on the new matrix from PCA. The new model will have higher predictive power as it addresses some issues regarding multi-collinearity and over-fitting. According to the two graphs below, the first three PCs were selected for the linear regression model.

$$Z_j = X_{Uj} = U_{1j} * X_1 + ... + U_{pj} * X_p$$

The j-th factorial axis uj is the eigenvector of XX^T corresponding to the second largest eigenvalue λj of XX^T.

j-th factorial variable



(a) heatmap for the top 10 PCs (b) variance explained by the top 10 PCs

Figure 3. PCA Analysis

Linear Regression with Lasso Regularization

$$Y = (X\beta)^T(Y - X\beta) + \lambda|\beta|_1 \tag{2}$$

Target variable cnt is log-transformed for linear regression assumptions, and the best λ chosen is 0.0023 by grid search and cross-validation with train to test ratio 80-20.

Result of feature importance is shown below

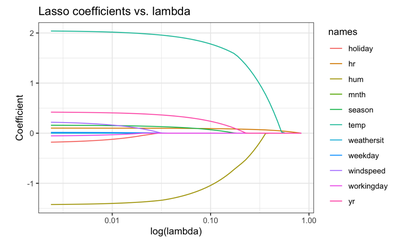


Figure 4. Lasso feature importance

Results

Model/Metrics	Simple LR	PCA LR	Lasso LR
R^2	0.471	**0.591**	0.478
MSE	1.157	**0.899**	1.190
MAE	0.843	**0.715**	0.852

Table 1. models metrics comparison

- **Champion model**: Linear model with PCA has the lowest MSE and highest R^2
  - $MSE \downarrow$, model accuracy ↑
  - $R^2 \uparrow$, proportion of variance explained ↑
  - $MAE \downarrow$, model accuracy ↑
- **Global Feature Importance**: AIC, BIC, Lasso, PCA
  - Both AIC and BIC indicate the most **important features** of the model are **season, temperature, holiday, humidity, hours, and years**.
  - LASSO and PCA both demonstrate important features same as AIC and BIC, and the **least important factor** demostrated by Lasso is **weekday**
- **Model Implication**
  - **General point**: People are more willing to ride a bike if the weather or temperature is comfortable enough.
  - **Surprising point**: factors indicating possible changes in population in a given area, such as working days and holidays, do not affect the bike rental.

References & Contribution

UCI Machine Learning Repo: Bike Sharing Dataset

https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset.

*Each person equally contributed to this project