Bike Sharing Demand Forecast

Junlin Lyu, Hongtu Zhang, Ziyu Zhou

Abstract

In this paper, our project is to predict the number of sharing bikes on a certain day, we used three kinds of regression models to make forecast. This paper applied several supervised learning regression models, including Linear Regression, Principal Component Analysis on Linear regression, Decision Tree, Random Forest and Feedforward Neural Networks on Bike Sharing dataset, which has 13 features. After considering R-squared and RMSE, the experimental results show that Feedforward Neural Networks outperform other models, which means that this model is suitable to analyze Sharing Bike problem.

1. Introduction

As one of the most important technical progress, the transportation method has played a dispensable role throughout the history of the human civilization. From walking to the livestock in the farming age, and to the various mechanical machines such as vehicles, airplanes and ships, the innovation of the transportation has always been connected with the development of the technology, which could effectively improve the operation of the society. Recently in the 21st Century, since existing transportation methods could not match the growing scale of the cities and the huge population base, a new concept of “Sharing Transportation” has been created and used widely in many countries. “Sharing Bike” is one of the most popular methods, which means that people could rent a bike from somewhere they need and return it at their destination.

However, till now, there are many aspects to be optimized, such as how to decide the number of bikes placed in the city to lower the cost and avoid wasting resource. We propose to use supervised machine learning skills to optimize “Sharing Bike” amount to reduce the intense transportation pressure and satisfy people’s demands.

2. Technical Approach

Now that our project is to predict the number of sharing bikes on a certain day, we plan to use three kinds of regression models to make forecast. Firstly, we try to use the linear regression model and compare with the result after using Principal Component Analysis (PCA). We also use ridge and lasso regression to reduce model complexity and prevent overfitting which may result from simple linear regression. Regression Tree model is the second method, applying tree map to make prediction. Then we use random forest to make prediction by finding the optimal hyperparameters. The last method is Neural Networks, which is used to make prediction by setting specific activation function and number of hidden layers and nodes. After using three regression methods, we apply the R-squared and root mean square error (RMSE) to evaluate those models’ performance.

3. Experimental Results

3.1 Dataset

We use the “hour.csv” dataset [1] from UCI Machine Learning Repository. There are 17379 records corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA. Table 1 shows all important features.

Name	Type	Introduction
season	Integer	1: Springer, 2: Summer, 3: Fall, 4: Winter
yr	Integer	Year (0: 2011, 1:2012)
mnth	Integer	Month (1 to 12)
hr	Integer	Hour (0 to 23)
holiday	Integer	0: No holiday, 1: Holiday
weekday	Integer	Day of the week
workingday	Integer	0: Weekend or holiday 1: Neither weekend nor holiday
weathersit	Integer	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	Float	Temperature in Celsius
atemp	Float	Feeling temperature in Celsius
hum	Float	Humidity
windspeed	Float	Wind speed
cnt	Integer	Count of total rental bikes including both casual and registered

Table 1: Feature Description

3.2 Data Preprocessing

Firstly, we split “dteday” into “Year” and “day”. By making correlation analysis, we find that “season” is highly correlated with “mnth’ and “atemp” is highly correlated with “temp”, so we choose to drop “season” and “atemp”. Then we use Three-Sigma Limit to remove 244 outliers and keep 17135 samples for further modeling. After that, we create dummy variables for 7 features: “Year”, “mnth”, “hr”, “holiday”, “weekday”, “workingday” and “weathersit”. Lastly, we standardize the data and randomly split the dataset into training data (70%) and testing data (30%).

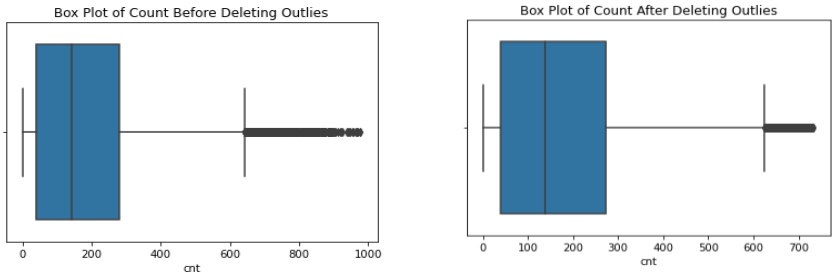

Figure 1: Box plot changes of deleting outliers

3.3 Data Visualization

Figure 2 shows that there are fewer demands in winter and holidays and on weekends. At the commuter time, from 7 to 8 o'clock and from 17 to 19 o'clock, the demands for bike rental are higher.

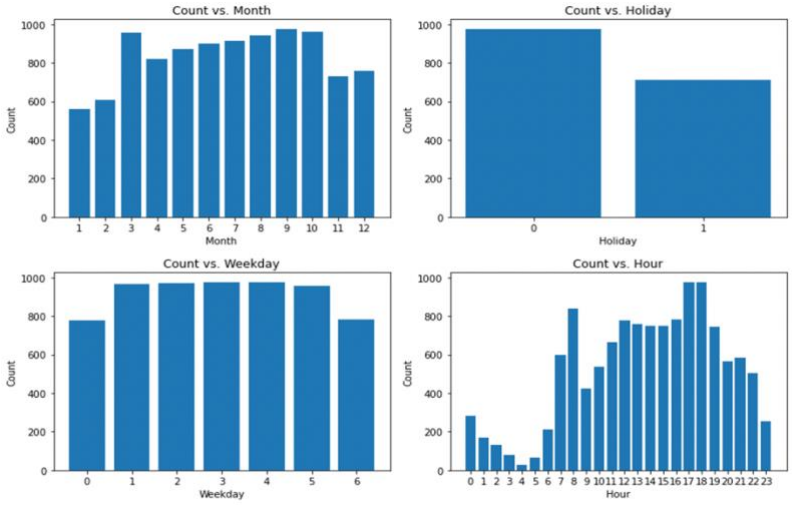

Figure 2: Count vs. Month, Holiday, Weekday, Hour

There are fewer demands under worse weather as the Figure 3 shows. When the humidity is extremely low, the amount of bike rental is small. The amount of bike rental has a negative relationship with the windspeed. Besides, people prefer to rent bikes when it's warm.

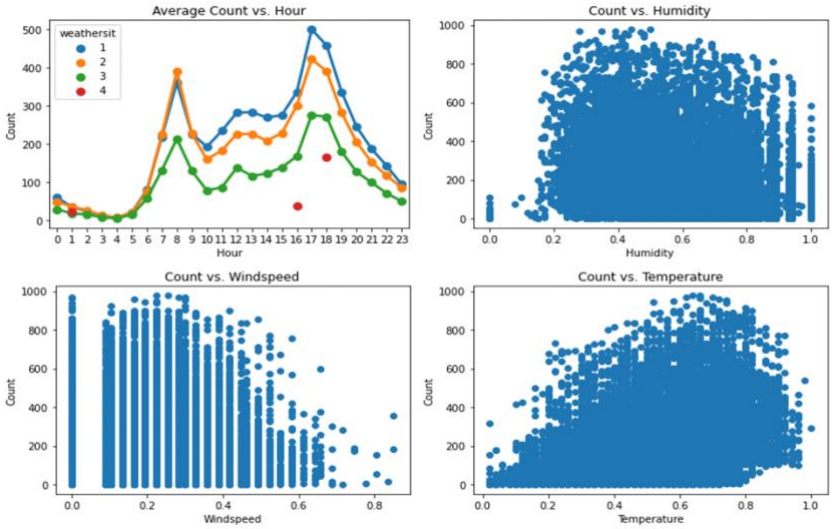

Figure 3: Count vs. Hour, Humidity, Windspeed, Temperature

3.4 Implementation

3.4.1 Linear Regression

First, we make use of linear regression model to make prediction and get an extremely large RMSE. One possible reason for the large RMSE is the curse of dimensionality as we have 54 features in all after transforming categorical variables into dummy variables. Then we use two approaches to solve the problem: one is using regularization to reduce overfitting (or underfitting); the second is to do principal component analysis (PCA) to achieve dimension reduction. We set n_components equal to 0.95 to keep 95% variance or information of the data. So, we get 6 linear regression models: Linear regression with PCA, Ridge Regression (L2 Norm) with PCA, Lasso Regression (L1 norm) with PCA, Linear regression without PCA, Ridge Regression (L2 Norm) without PCA and Lasso Regression (L1 norm) without PCA. After modeling and comparison, we find Ridge Regression (L2 norm) without PCA achieve lowest RMSE 95.6 (Table 2).

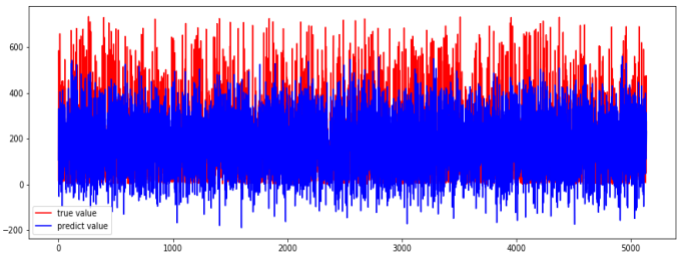

Figure 4: True Value vs. Predicted Value

Without PCA	RMSE	R2_score
Linear Regression	5.87e+13	-1.23e+23
Ridge Regression (L2 Norm)	95.6	0.67
Lasso Regression (L1 norm)	95.88	0.67
With PCA	**RMSE**	**R2_score**
Linear Regression	101.72	0.63
Ridge Regression (L2 Norm)	101.72	0.63
Lasso Regression (L1 norm)	101.92	0.63

Table 2: Linear Regression Model Comparison

3.4.2 Regression Tree

We use decision tree to build regression models in the form of a tree structure. Figure 5 shows the relationship between mean squared error and maximum depth. The MSE is negatively related with the maximum depth. When the depth over 10, there are no obvious decreases on the MSE. In order to reduce the time of machine learning we choose 10 as our maximum depth. The RMSE of this model is 95.81 and the r2_score is 0.67. From the Figure 6, we can see that the decision tree model does not fit well when the true value is large.

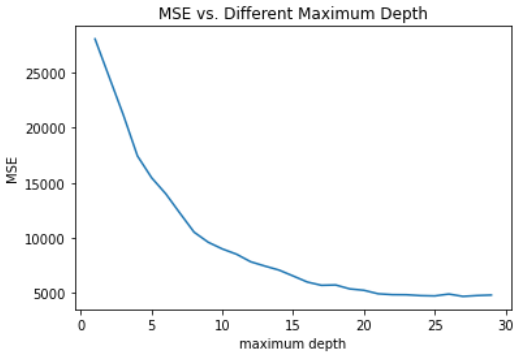

Figure 5: MSE vs. Maximum Depth

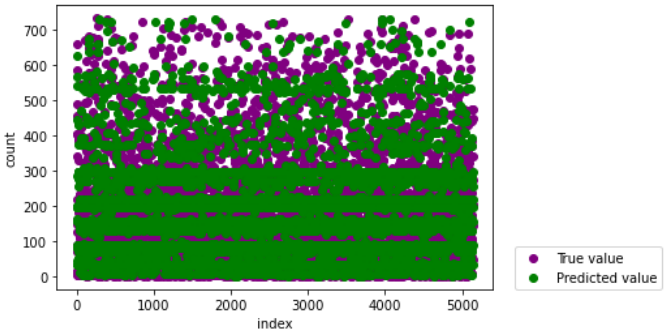

Figure 6: True Value vs. Predicted Value (Decision Tree)

Then we use random forest which is used to create an uncorrelated forest of trees whose prediction is more accurate than that of any individual trees. We have tried different hyperparameter combinations to find the best parameters. When the max_depth is 20 and n_estimators is 200, the model performs best. The r2_score of random forest regression is 0.89, indicating that the model explains most of the variability of the response data around its mean. Compared with the decision tree model, the RMSE of the random forest model

decreases to 54.89, achieving a more accurate prediction as Figure 7 shows. From table 3, we can get that the temperature, humidity and commuter time have much more significant influence on our prediction.

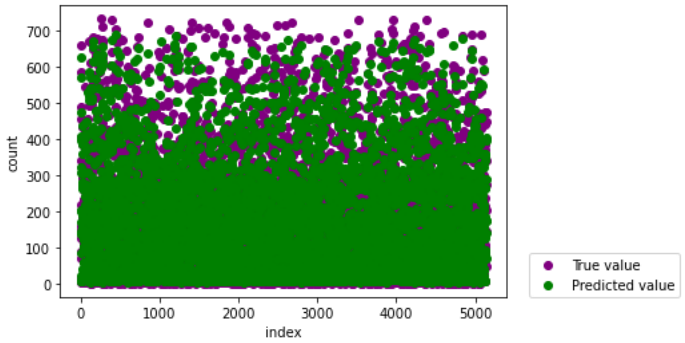

Figure 7: True Value vs. Predicted Value (Random Forest)

Features	Importance
temp	0.167348
hum	0.152843
hr_8	0.062749
hr_17	0.055790
hr_18	0.050134

Table 3: Top 5 Importance Features

3.4.3 Feedforward Neural Networks

The feedforward neural network is applied to make prediction of the amount of sharing bikes. As Figure 8 shows, the fully collective neural network has 2 hidden layers with 10 nodes for each hidden layer. The rectified linear activation function (ReLU) is used as the activation function of the model.

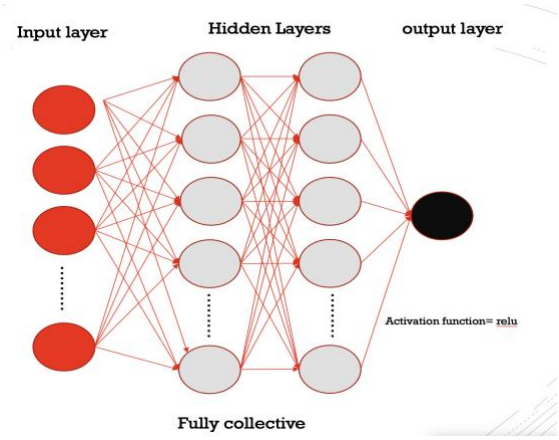

Figure 8: Neural Network Architecture

After building the model, we also optimize the model. We define SSE as the loss function and SGD as the optimizer. Finally, the RMSE of neural networks is 46.85 and the r2_score is 0.92.

To be clearer about how the model fits the new data, we make a comparison between the predicted value and the actual value. From Table 4, it is clear to see that the predicted value is pretty close to the actual value, which means the model could fit new data and perform well.

	Actual count	Predicted count
0	17.0	9.466507
1	584.0	532.974060
2	404.0	407.823242
3	457.0	423.696167
4	127.0	133.015915
...	...	...
5136	447.0	526.647034
5137	343.0	320.252441
5138	474.0	467.928802
5139	146.0	141.205750
5140	274.0	191.438995

[5141 rows x 2 columns]

Table 4: Comparison between actual values and predicted values

3.5 Conclusion

The paper presents three kinds of regression models of sharing bikes, Linear regression, Regression tree and Feedforward Neural Networks. For Linear regression, without PCA or regularization, the result seems less successful and the RMSE is large enough. While we consider using regularization and PCA, it seems much better for fitting the testing data. But still, the r2_score of the best linear regression model (Ridge regression without PCA) is 0.67, not that successful for analysis and prediction. We need more data for training if we choose to use Linear regression as our model selection. Then we consider the other two model analysis. One is Regression tree. This model is quiet well for analysis and prediction. We find there are three elements that are most important for our analysis: temperature, humidity and commuter time. On top of that, Feedforward Neural Network with 2 hidden layers and 10 nodes for each hidden layer performs best. The neural network achieves smallest RMSE and largest r2_score compared with other models.

Model	RMSE	R2_score
Ridge Regression (L2 Norm)	95.6	0.67
Random Forest Regression	54.89	0.89
Feedforward Neural Networks	46.85	0.92

Table 5: Comparison between different models

Therefore, companies that run the business of sharing bikes can rely on the neural network model to make forecast. They are supposed to adjust the supply of bikes according to the weather condition and increase supply during the commuter time to satisfy the market's demands.

4. Participants Contribution

The division of work throughout the project is evenly distributed between authors, although the principal tasks not the same. Mr. Zhang is responsible for data preprocessing. The linear regression model is made by Miss. Zhou; the regression tree model is made by Mr. Zhang; the feedforward neural network model is made by Mr. Lyu. The report is completed in cooperation.

References

[1] Hadi Fanaee-T. *Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto,* archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset.