Study on Forecasting Public Revenue for Andhra Pradesh State Road Transport Using Regression Techniques

Reshma Gummadi¹*, Sreenivasa Reddy Edara²

¹Research Scholar, Acharya Nagarjuna University, Guntur, Andhra Pradesh, 522510
²Department of Computer Science & Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, 522510
*Corresponding author E-mail: reshma.gorripati@gmail.com

Abstract

Forecasting public revenue of transport system indirectly shows the development of system with the growth of economy. Andhra Pradesh state road transport corporation (APSRTC) develops rapidly by this process. Therefore there is a need for developing model to forecast public revenue and identifying the relationship between transportation and economy of the system. The passenger traffic volume collected from the APSRTC for calculating public Revenue is based on the data from 2016 to 2017. Then different regression models are applied on the collected data for analysis purpose. The linear regression outperforms gradient boosting, decision trees and neural networks with a relative error less than 5% for predicting public revenue. The accuracy of models on revenue has been improved which helps for the development of transportation system.

Keywords: Public Revenue; Linear Regression; Gradient Boosting.

1. Introduction

Forecasting public revenue based on number of passengers traveling on the bus helps for planning and development of transport system. Different forecast approaches including regression techniques applied on the data which helps in discovering the relationships among the variables [2]. There are various regression techniques which help to forecast accurately. Regression techniques include linear regression, gradient boosting regression, and random forest regression and so on. In this paper general procedure for gradient booting algorithm and linear regression is explained and implemented in python to know the predicted values of public revenue. By plotting graphs linear regression and gradient boosting regression techniques were compared. The relationship between Income and passenger flow is difficult to analyze. This paper forecast the revenue using regression techniques to improve accuracy. Linear regression gave better forecasting results compared to Random forest and gradient boosting.

2. Forecasting revenue using regression techniques

2.1. Data source

The Transit bus information of route wise data in this case from macherla is extracted from databases for the period of April 2016 to 31st December 2016. To forecast the income firstly the ticket fare of passengers paid on each date is calculated. So before applying regression techniques the total amount on that date paid by all passengers travelled is ready.

Dataset = data.groupby(['Date'],as_index=False)[['TotalAmount']].sum()
# declaration of temporary arrays to store the data.
x_data = []
y_data = []

for d in range(6,dataset.shape[0]):
    print(d)
# The matrix will have columns one for each moth, prior to the one to be predicted, here start with the seventh month available.
x = dataset.iloc[d-6:d].values.ravel()
y = dataset.iloc[d].values[0]
x_data.append(x)
y_data.append(y)
x_data = np.array(x_data)
y_data = np.array(y_data)

Print (y_data)
Gradient Boosting is a special type of boosting technique which reduces errors sequentially. In Boosting, Models are built in series. In each model, the heights are adjusted based on learning from previous models. After applying 1st and second model, the errors are identified and reduced in 3rd model. The term boosting itself means combining weak models to create powerful models. Decision tree is one of the most popular choices of ensemble models.

2.2.1. General procedure for gradient boosting
1) Develop a regression tree. For any leaf node, it gives an average value.
2) Find errors for all the observations of leaf nodes.
3) Consider the above errors as new observation values.
4) Develop second Tree to minimize these errors.
5) Repeat steps 3 to 4 for a specified number of iterations.

2.2.2. Gradient boosting in python
The data collected from APSRTC consists of number of fields. For this purpose the date is grouped from entire data and sum of the amount paid by the passengers are taken in to consideration.

```python
# loading a regression predictor can be done by the following steps.
for i in range(30,end):
x_tr = x_data[:i, :]
y_tr = y_data[:i]
x_ts = x_data[i, :]
y_ts = y_data[i]
model = LinearRegression()
model.fit(x_tr, y_tr)
y_pred.append(model.predict(x_ts))
y_pred_last.append(x_ts[-1])
y_pred_ma.append(x_test.mean())
y_true.append(y_test)
```

# computing error residuals.
```python
aa = list(y_pred)
bbs = [list(x) for x in aa]
y_pred_1 = []
for i in bb:
    y_pred_1.append(i[0])
```

2.3. Implementation in python
After plotting the graph the view is shown below in Figure 2.
3. Forecasting revenue based on linear regression

Linear regression is a linear model which a linear relationship can be estimated between input variables and single output variable when the input variable is a single variable then that model is called simple linear regression. The input variable is called independent variable and the output variable is called dependent variable [5]. The linear relationship can be defined as

\[ Y = \beta_0 + \beta_1 x \]

Where \( \beta_1 \) is called coefficient and \( \beta_0 \) is called bias coefficient. The coefficients can be found using the ordinary least square method and gradient descent approach. The ordinary least square method is used in simple linear regression, a good model will always have minimum error. A total error of this model is the sum of all errors of each point [5].

\[ D = \sum d_i^2 \]

\( D_i \) is the distance between line and where \( m \) is the total number of points. The model can be evaluated using Root Mean squared Error and coefficient of determination. RMSE is the square root of sum of all errors divided by number of values[5].

3.1. Implementation in python

1) Linear regression cannot be applied to the entire dataset; hence the data is splitted in to training data and test data. Now model can be trained on training data and testing can be performed on test data. The linear regression model can be fitted to the training data set.

```python
y_pred_last1 = []
y_pred_last1.append(x_test[-1])
y_pred_last1.append(x_test.mean())
y_true1.append(y_test)
model.predict([166621, 175371, 163806, 153404, 180182, 156400])
print(model.predict(x_test))

[168395.69898104 163972.35456069 167979.48902895 168395.69898104 166158.12370387 168395.69898104 173082.38818986]
```

2) Creating a difference transform of the dataset and make a prediction which give regression coefficients and lags

```python
def difference(dataset):
diff = list()
for i in range(1, len(dataset)):
    value = dataset[i] - dataset[i-1]
diff.append(value)
return numpy.array(diff)

def predict1(coef, history):
y = coef[0]
for j in range(1, len(coef)):
y += coef[i] * history[-j]
return y
```

3) Split the dataset and train the data using auto regression. While during testing walk forward over time steps. The test MSE is obtained and plotted figure obtained as shown in Figure 3.

Test MSE: 938407405.243

4. Results and discussion

The following table shows the results obtained for various regression techniques applied on the given data. The actual test values are taken from the month of December 23rd to 31st. linear regression technique goes well rather gradient boosting technique. The public revenue of the passengers also increases with respect to time more than 5%.
5. Conclusion

The methods used for forecasting public revenue based on passenger data mainly focused on various regression techniques. In this paper we use linear regression and gradient boosting technique for forecasting. Linear regression gives better predicted values compared to gradient boosting technique. There is still need for refining the data and to achieve more accurate results for the development of transportation systems. The other techniques like random forest algorithm and neural networks can be applied and compared for better accurate results. The predicted data is in terms of revenue per year and by calculating the more accurate the data the prediction can be done effectively which helps in the analysis of the revenue of public transport system.

References


