Time Series Analysis to Forecast Traffic

41202, 41203, 41205: Aditi Wikhe, Jagruti Agrawal,
Anuraag Shankar

Time series regression is a statistical method for predicting a future response based on the response history (known as autoregressive dynamics) and the transfer of dynamics from relevant predictors. Time series regression can help you understand and predict the behavior of dynamic systems from experimental or observational data. Common uses of time series regression include modeling and forecasting of economic, financial, biological, and engineering systems. In this project, we use Python and its libraries to perform the task of regression on a time series dataset to forecast traffic.

1 INTRODUCTION

A regression problem is when the output variable is a real or continuous value, such as “salary” or “weight”. Many different models can be used, the simplest is linear regression. It tries to fit data with the best hyperplane which goes through the points.

Regression Analysis is a statistical process for estimating the relationships between the dependent variables or criterion variables and one or more independent variables or predictors. Regression analysis explains the changes in criteria in relation to changes in select predictors. The conditional expectation of the criteria is based on predictors where the average value of the dependent variables is given when the independent variables are changed. Three major uses for regression analysis are

determining the strength of predictors, forecasting an effect, and trend forecasting.

A few common types of regression are:

1. **Linear regression:**
    a. Used for predictive analysis.
    b. Is a linear approach for modelling the relationship between the criterion or the scalar response and the multiple predictors or explanatory variables.
    c. Focuses on the conditional probability distribution of the response given the values of the predictors.
    d. For linear regression, there is a danger of overfitting.
    e. The formula for linear regression is: Y' = bX + A.
2. **Ridge regression:**
    a. Is a technique for analyzing multiple regression data.
    b. When multicollinearity occurs, least squares estimates are unbiased.
    c. A degree of bias is added to the regression estimates, and as a result, ridge regression reduces the standard errors.
3. **Lasso regression:**
    a. Is a regression analysis method that performs both variable selection and regularization.
    b. Lasso regression uses soft thresholding.
    c. Lasso regression selects only a subset of the provided covariates for use in the final model.

2 PROJECT SCOPE

Regression analysis in business is a statistical method used to find the relations between two or more independent and dependent variables. One variable is independent and its impact on the other dependent variables is measured. There are more than 10 types of regression models. When there is only one dependent and independent variable, or predictor variable, we call is simple regression. On the other hand, when there are many independent variables influencing one dependent variable, we call it multiple regression.

Linear regression model is a linear approach to modeling the relationship between a scalar response and one or many explanatory variables. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

Nonlinear regression is a form of regression analysis where data fits a model and is then expressed as a mathematical function. Nonlinear regression is computed by finding the difference between the fitted nonlinear function and every Y point of data in the set.

Regression has high usage in the business industry. A few examples are as follows:

1. **Predictive analytics:** Forecasting future opportunities and risks is the most prominent application of regression analysis in business.
2. **Operation Efficiency:** Regression models can also be used to optimize business processes.
3. **Supporting Decisions:** By reducing the tremendous amount of raw data into actionable information, regression analysis leads the way to smarter and more accurate decisions.
4. **Correcting Errors:** Regression is not only great for lending empirical support to management decisions but also for identifying errors in judgment.
5. **New Insights:** Over time businesses have gathered a large volume of unorganized data that has the potential to yield valuable insights.

3 REQUIREMENTS

The project has been built using Python on Google Colaboratory. The following libraries were used while building the project:

1. Pandas – pandas
2. NumPy – numpy
3. Sci-kit Learn – sklearn

The configurations of the notebook are as follows:

1. Processor Name: Intel(R) Xeon(R)
2. Processor Speed: 2.20 GHz
3. Operating System: Ubuntu 18.04 64-bit
4. RAM: 12.0 GB
5. GPU: Nvidia Tesla K80.

4 METHODOLOGY

The process of forecasting traffic on nodes consisted of several stages that are described below.

The initial dataset looked as follows:

ID	Datetime	Count
0	25-08-2012 00:00	8
1	25-08-2012 01:00	2
2	25-08-2012 02:00	6
3	25-08-2012 03:00	2
4	25-08-2012 04:00	2

Table 1. Initial dataset

The “Datetime” column was then split into two columns namely “Date” and “Time”. The head of the table is shown below:

ID	Date	Time	Count
0	25-08-2012	00:00	8
1	25-08-2012	01:00	2
2	25-08-2012	02:00	6
3	25-08-2012	03:00	2
4	25-08-2012	04:00	2

Table 2. Dataset after splitting the Datetime column

The next step was to convert the Date and Time columns into appropriate formats. For this, the following processing steps were applied:

1. Date column was given continuous values. The first date was given a value 0, second day 1 and so on.
2. Time was discretized i.e. each hour was given a value from 0-23 depending on what hour of the day it was.

The processed looked as follows:

ID	Date	Time	Count
0	0	0	8
1	0	1	2
2	0	2	6
3	0	3	2
4	0	4	2
…	…	…	…
18283	761	19	868
18284	761	20	732
18285	761	21	702
18286	761	22	580
18287	761	23	534

Table 3. Table after final processing step.

This data was fed into regression models. The count column acted as the target variable to predict.

5 RESULTS

The following three regression models were tested:

1. Linear
2. Ridge (alpha = 0.5)
3. Lasso (alpha = 0.5)

The dataset was split into training and testing in the ratio of 90:10.

The metrics used are Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Their formula is as follows:

$$MSE = (y_{true} - y_{pred})^2$$

$$RMSE = \sqrt{(y_{true} - y_{pred})^2}$$

The results obtained on the test set were as follows:

Regression Model	**MSE Loss**	**RMSE Loss**
Linear	8003.7	89.463
Ridge	8005.9	89.475
Lasso	8001.2	89.449

Table 4. Metrics obtained by the models

6 Conclusion

In this project we looked at a time series dataset and predicted traffic on its nodes. This was achieved by transforming the dataset into a format that the model could interpret. This was done using a series of processing steps which finally allowed us to feed the data into a regression model. Three different regression models were tested, and it can be observed that the Lasso model obtained the best results. But the metrics obtained by each model were really close to each other and there was no model which was far superior than the others.