BIG DATA FOR INDIAN RAILWAYS

Indrajit Bhosale, Yash Gandhi, Pune Institute of Computer Technology,Pune

Abstract:

Two concepts currently leading the Information Technology revolution are Big Data and Analytics. The public transportation industry has not been at the forefront in utilizing and implementing Analytics and Big Data on an extensive scale in developing countries like India. Our study gives an overview of Big Data technologies in context of transportation with specific to Railways. The Railway industry is an infrastructure intensive industry that relies on large scale of data to operate, maintain, guarantee the safety of railway systems and control their cost. Our study also gives methodology on how the existing data modules from the transport authority can be combined with Big Data algorithms in providing predictive maintenance decision making. Rails are systematically inspected for defects using various non-destructive inspection and evaluation techniques with help of sensors. The data obtained by these sensors is increasing in both quality and quantity and is helpful for a proper time series analysis of the trends and changes in the Rail system over due course of time. Also, this study evaluates train timetable from the viewpoint of Big Data on rail transit lines. By calculating the load factor of rail lines for different times and zones, the proposed model assesses the drawbacks of scheduled timetables. By starting to make extensive use of Railway's “Big Data” , we propose optimize its capital infrastructure and safely manage its operations while keeping costs under control with comparisons of historic cost data and optimized cost after making proper use of the proposed system. We conduct this study with parallel references from The Dutch and American Railways and try to replicate/model them for Indian Railways.

Keywords: Big Data for Railways, Predictive Algorithms, Sensor Data, Maintenance Decisions

I. INTRODUCTION

Problem Statement

Determine problems faced in Indian Railways where the concept of Big Data can be implemented and come up with possible solutions.

Motivation

Indian Railways is one of the biggest employer in the world. Just to put things in perspective Indian Railways transports whole population of Australia daily, such is the size of Indian Railways. Such a large system generates large amount of data and hence can be easily conferred as big data. Big Data analytics and other analysis have now started to be used in western countries. We had a long standing desire to do something for the Digital India initiative. What better sector to contribute than railways. We wanted to implement big data algorithms on the railway system. And luckily to our benefit we found complete dataset for Indian Railways on data.gov.in. This was motivation enough for us and we decided to take this as our research topic and potential project topic

SCOPE

There are high prospects for using the Indian Railways' available data and organizing, optimizing, improvising and identifying hidden patterns in it to make better decisions for best and effective Railway System. Big Data Analytics comprises of algorithms that can filter out unnecessary data and generate value from the existing data and also help in making better predictive decisions. Indian Railway network being amongst the biggest railway networks in the World , such a unique combination of Big Data and Railways has been tested and implemented for French, Dutch, USA and Chinese Railways and can be modeled for Indian Railways too.

II. LITERATURE SURVEY

CURRENT STATE OF RAILWAYS

Predictive Maintenance:

Current maintenance of railroads in India:-

(1) System to be adopted- The track should be maintained either by conventional system of track maintenance or by 3-tier system of track maintenance.

(2) Details Of Maintenance Works- (a) In both the systems, track requires to be overhauled periodically with the object of restoring it to best possible condition, consistent with its maintainability. Periodicity of overhauling depends on several factors, such as type of track structure, its age, maximum permissible speed, system of traction, condition of formation etc. Irrespective of the system of track maintenance adopted, it is obligatory to overhaul specified lengths of gang beat annually. The length of the section to be overhauled shall be such that complete overhauling of track will be accomplished within a specific period (normally 3 to 5 years).

[1].

Source Indian Railways Website.

New Model

Classify Stations, Tracks based on load factor, volume, train density using ml and data mining classification algorithms.

K-means,One vs All, Naive-Bayes

Develop a model with stations sorted according to priority of maintenance.

Financial State

It is important that a continuous and concurrent watch is kept on the realisation of earnings as envisaged in the Budget. This is done through the medium of a tendaystatement of earnings on "originating" basis, the statement for the last period of the monthgiving also the position for the month and the cumulative position from 1st April to end of themonth. These statements should give also the proportionate budgeted earning on originatingbasis and the actuals for and to end of the relevant period of the preceding year for comparison.The originating basis is adopted to secure prompt reporting since the Railway wise apportioned
earnings for each month do not become available until a few weeks later.
Investment decisions are among the most interesting and difficult decisions to be made by the Managements. It is fundamental to railway system as a commercial undertaking that expenditure other than that wholly chargeable to Ordinary Revenue incurred on new assets or for improvement of existing assets should be financially justified and sanctioned before it is actually incurred.

As an exception to above mentioned stuff, while no financial justification as such need be given in the
following cases, it should be seen that the scale of expenditure incurred is as economical as possible
consistent with the extant orders, if any, on the subject:-
(a) when the expenditure is incurred on a statutory obligation (for example, the fencing of
machinery) ;
(b) when the expenditure is unavoidable on considerations of safety;
(c) when the expenditure is incurred on passenger amenity works; and
(d) When the expenditure is incurred on labour welfare works except residential
buildings for which special rules are applicable.

Interest during construction should be added to the cost (excluding that chargeable to Revenue)of the projects, the construction of which is likely to last for more than one year. In the case of construction of bridges, maintenance charges should include, besides the maintenance charges on the bridges proper, the maintenance charges of the training works also.

For the purpose of carrying out a meaningful comparison of the actual working expenses for (and to end of) the month with the budget allotment, it is necessary to distribute the sanctioned allotment for the year over the twelvemonths after taking all known factors of disturbance or special features into account. While theres posibility for the control of expenditure against the budget allotment devolves upon the authority at whose disposal the allotment has been placed, It is the duty of the Accounts Officer,in his capacity as the financial adviser to the Administration, to render all possible assistance to the controlling authorities in the exercise of such control. Accordingly, he works out, at the beginning of each financial year, in consultation with the officers responsible for the control of expenditure, the estimated progressive expenditure under each sub head of a grant keeping in view the following factors:
(i) Throw forward from the previous year.
(ii) All expenditure whether in cash or by transfer, the liability for which already exists,but which is not likely to be distributed evenly during the year, whether because it is of a periodical nature, or because it is contingent on the receipt of supplies, or for any other reason.
(iii) Expenditure which is practically fixed and evenly distributed throughout the year.
(iv) Other expenditure which is likely to be incurred during the year but liabilities for
which have yet to be incurred.
(v) The need to keep some amount as a reserve for meeting fresh or unanticipated
expenditure.

It may sometimes be necessary to reject a more economical alternative, because of considerations on which it is difficult to put a precise money value. If on the strength of any such factor a proposal is adopted that is less economical than its alternatives, the reasons determining such a choice should be recorded in the report of the estimate containing such proposals.

The earnings expected from additional traffic whether on the existing line or on a new line should be very carefully estimated in the traffic survey report. The estimate of earnings should be worked out for each commodity and for the lead from its origin up to the point of termination of the traffic. It should be ensured, however, that in case line capacity works are required to be undertaken in the contiguous section or Railway system over which the additional traffic is expected to be moved, the initial cost estimate for the scheme takes into account the cost of such additional line capacity works over the entire load of the traffic.

ISSUES FACED

Issues in Financial System

The majority of the new projects related to the railway will have to be funded by public-private partnerships, according to Suresh Prabhu, the Indian Railway Minister. He said the government is relying on foreign investment to fund station upgrades and new bullet trains. The cabinet will approve FDI in railway infrastructure, which has previously not been allowed.

The move comes in acknowledgement of the fact that India’s current fiscal issues present a significant problem to major investments such as this. As Pradhu remarked: “Internal resources are insufficient to meet the requirement”.

Given the magnitude of the railway upgrade, it is no question that the investment needed will be of the long-term type, with realistic estimates suggesting it will take a minimum of 10 years for the network to reach modern standards. In order to not be overwhelmed and repeat the mistakes of past governments, Prabhu has taken a pragmatic approach that divides the upgrade into several smaller projects that will be tackled one at a time. These will include the establishment of the high-speed rail network and increasing speed on the current structure – the two projects considered top priorities so far.

This division of projects is crucial, says Prabhu, pointing to how previous PPPs have failed to attract sufficient investment.

Concerns about risk are a particular hurdle for India in its attempts to attract FDI, but Agarwal maintains investors must keep their eye on the risk-return equation instead. Another issue pertains to the lack of long-term financing in India for projects such as these, as well as the governments tradition of bundling infrastructure projects in with real estate. It is crucial that the upgrade doesn't depend on such multi-projects, as the risk appetite of infrastructure investors does not match up to that of real estate and could prevent the inflow of FDI. The government and the Reserve Bank of India will also have to work hard at reforming banks lending practices – and this, again, could prove a long-term affair.

An independent railway regulator is necessary to deal with economic side of the railways and de-politicise issues such as tariffs and prioritisation of projects," says Agarwal. "It's taken 10 years for the government to do a 14 percent rate increase, so government involvement needs to be reduced. The regulator should be able to gauge investments and tariffs based on demand and costs, and then the government could subsidise projects or passenger classes in a more targeted manner."

To this end, the Indian Transport Minister has proposed establishing automatic revisions to the subsidised passenger fares in order to reduce political skirmishing over increases needed to cover rising costs. For decades, successive Indian governments have held passenger fares far below their costs to keep trains universally affordable, while charging steep freight rates in order to cover the losses. This has hurt commercial train freight in India, but is not by any means the only issue keeping companies on the roads and away from the tracks.

Analysis of Predictive Models for Maintenance

ALGORITHM 1:

K-MEANS clustering intends to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function.

Algorithm

Clusters the data into k groups where k is predefined.

Select k points at random as cluster centers.

Assign objects to their closest cluster center according to the Euclidean distance function.

Calculate the centroid or mean of all objects in each cluster.

K-Means is relatively an efficient method. However, we need to specify the number of clusters, in advance and the final results are sensitive to initialization and often terminates at a local optimum. There is no global theoretical method to find the optimal number of clusters. A practical approach is to compare the outcomes of multiple runs with different k and choose the best one based on a predefined criterion. In general, a large k probably decreases the error but increases the risk .

Example:

Suppose we want to group the visitors to a website using just their age (a one-dimensional space) as follows:

15,15,16,19,19,20,20,21,22,28,35,40,41,42,43,44,60,61,65

Initial clusters:

Centroid (C1) = 16 [1]

Centroid (C2) = 22 [2]

Iteration 1:

C1 = 15.33 [15,15,16]

C2 = 36.25 [19,19,20,20,21,22,28,35,4,41,42,43,4,60,61,65]

Iteration 2:

C1 = 18.56 [15,15,16,19,19,20,20,21,22]

C2 = 45.90 [2,35,40,41,4243,4,60,61,65]

Iteration 3:

C1 = 19.50 [15,15,16,19,19,20,20,21,22,28]

C2 = 47.89 [35,40,41,42,43,44,60,61,65]

No change between iterations 3 and 4 has been noted. By using clustering, 2 groups have been identified 15-28 and 35-65. The initial choice of centroids can affect the output clusters, so the algorithm is often run multiple times with different starting conditions in order to get a fair view of what the clusters should be.

ALGORITHM 2 :

The NAIVE BAYES classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Algorithm

Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

P(c|x) is the posterior probability of class (target) given predictor (attribute).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor.

ALGORITHM 3:

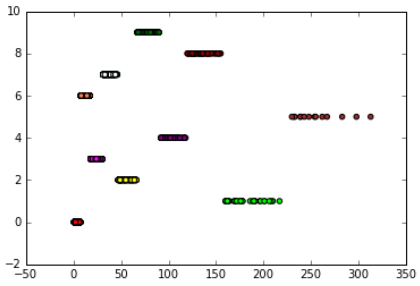
MULTICLASS LOGISTIC REGRESSION(One-vs.-rest):

The one-vs.-rest strategy involves training a single classifier per class, with the samples of that class as positive samples and all other samples as negatives. This strategy requires the base classifiers to produce a real-valued confidence score for its decision, rather than just a class label; discrete class labels alone can lead to ambiguities, where multiple classes are predicted for a single sample.

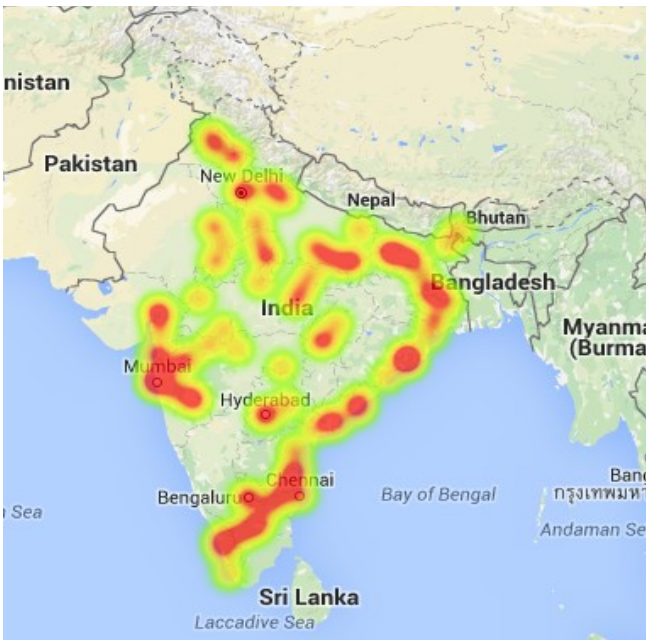
Implementation of proposed model for Indian Railway Time-Table

K-MEANS OUTPUT:-

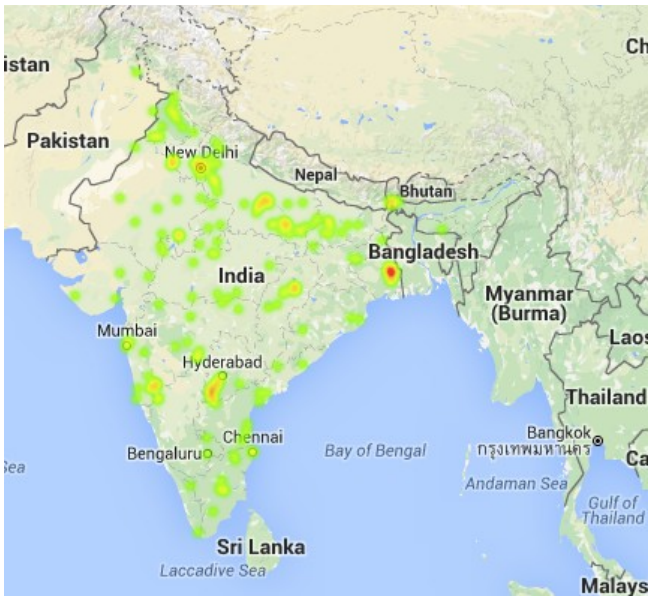
```
plt.scatter(x.load, model.labels_, c=colormap[model.labels_])
<matplotlib.collections.PathCollection at 0x7f495c291dd0>
```



HEAT-MAP FOR MAXIMUM PRIORITY STATIONS:-



HEAT-MAP FOR LEAST PRIORITY STATIONS:-



Analysis of Finances :

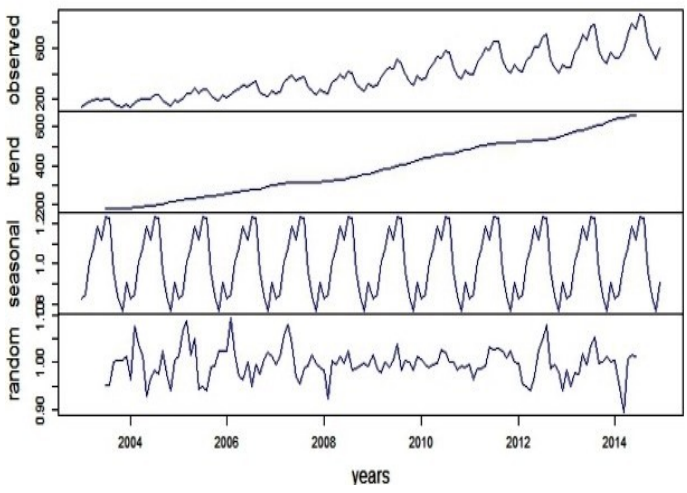
ARIMA Time Series Analysis Model for Finance

The fundamental idea for time series analysis is to decompose the original time series (sales, stock market trends, etc.) into several independent components. Typically, business time series are divided into the following four components:

- **Trend** – overall direction of the series i.e. upwards, downwards etc.
- **Seasonality** – monthly or quarterly patterns
- **Cycle** – long term business cycles
- **Irregular remainder** – random noise left after extraction of all the components

Interference of these components produces the final series.

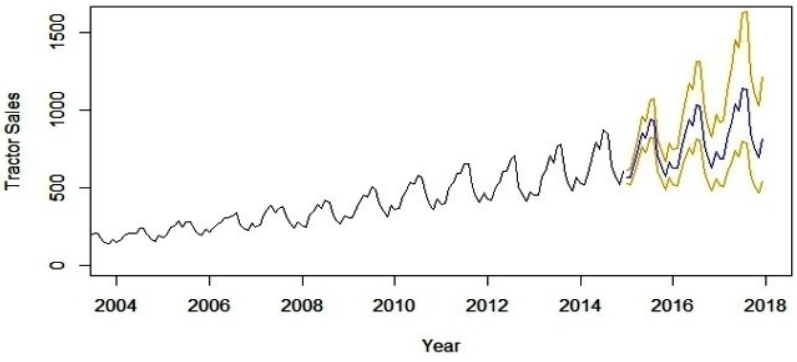
Now the question is: why bother decomposing the original / actual time series into components? The answer: It is much easier to forecast the individual regular patterns produced through decomposition of time series then the actual series. This is similar to reproduction and forecasting the individual sine waves (A, B, C, and D) instead of the final irregular pattern produced through the product of these four sine waves.



1st Pass of ARIMA :
Integrated (I) – subtract time series with its lagged series to extract trends from the data.

2nd Pass of ARIMA :
AutoRegressive (AR) – extract the influence of the previous periods' values on the current period.

3rd Pass of ARIMA:
Moving Average (MA) – extract the influence of the previous periods error terms on the current periods error



Forecasting Results

Conclusion

Big Data is the next frontier for both scientific and technical innovation in different parts of our society. Although big data is at the nascent stage, for railroad data both the quality and quantity are going to increase and this is the time for industries to make use of the available information and literature to develop industry-specific large data analytics approaches and methods. Successful completion of this agenda will put the railway industries in a new league through the development of advanced decision making safety and efficiency for railroads.

In today's era of system optimization and cost sensitivity, the proper planning and management of track maintenance can be an important tool in assisting railways in controlling their capital and maintenance costs. This is particularly true for high capital cost items such as rail, ties, and ballast which represents a sizable portion of a railway's maintenance of way budget. As the cost of maintenance continues to increase, the ability to properly plan and execute track maintenance programs in an efficient and cost- effective manner becomes increasingly important.

Future Scope

Although big data is at the starting stage for rail inspection, inspection data are going to increase in both quality and quantity. It is the time for railway industry to make full use of inspection data and advanced artificial intelligence technologies to develop new and efficient rail inspection platform, which should improve the safety and efficiency of railway systems.
Forecasting models are combined with the large data bases in maintenance planning models for determination of maintenance requirements and scheduling of maintenance activities across these large networks. The increasing use of inspection technology and the concurrent introduction of new technologies and associated analyses techniques provide railways with increasingly accurate and timely information about the condition of the track and its key components. This inspection data, which provides the railway with detailed information about the condition of the rail, forms the basis for improved planning, analysis and management based on the actual condition of the track. This use of condition based planning and management has been shown to be a significantly more efficient method of managing the rail asset than the traditional "rules" based approach, because it takes into account the local differences in behavior and performance, as they effect the degradation of the rails. In addition, it allows for a more accurate planning and scheduling of rail maintenance activities, since the times and locations for the key production activities are more accurately known.

REFERENCES

[1] References :
[2]
[3] a. Evaluating rail transit timetable using big passengers' data ,
[4] Zhibin Jiang, Ching-Hsien Hsu, Daqiang Zhang, Xiaolei Zou ,
[5] To appear in: Journal of Computer and System Sciences(Science Direct)
[6]
[7] b. Rail Inspection Meets Big Data: Methods and Trends ,
[8] Qingyong Li, Zhangdui Zhong, Zhengping Liang and Yong Liang ,
[9] 2015 18th International Conference on Network-Based Information Systems
[10]
[11] c. Railway Assets: A Potential Domain for Big Data Analytics ,
[12] Adithya Thaduri , Diego Galar , Uday Kumar ,
[13] INNS Conference on Big Data 2015 Program San Francisco, CA, USA
[14] P:F-SMR-UG / 02 / R0
[15]
[16] d. Big Data Challenges in Railway Engineering ,
[17] Nii Attoh-Okine , 2014 IEEE International Conference on Big Data
[18]
[19] e. Some Examples of Big Data in Railroad Engineering ,
[20] Allan M Zarembski , 2014 IEEE International Conference on Big Data