Air Traffic Control using Big Data Analysis and Machine Learning

Preeti Verma[1], Anusha Ramachandran[2], Prachi Dhariwal[3] and Parul Yadav[4]

Student-BVCOE, Student-BVCOE, Student-BVCOE, Assistant Prof.-BVCOE
Information Technology Department, Bharati Vidyapeeth's College of Engineering
Guru Gobind Singh Indraprastha University
New Delhi, India

ABSTRACT - *Nowadays, as there is a rapid increase in the amount of air traffic, so does the compulsion for virtuous, globally correlated and interoperable Air Traffic Management systems. We are searching for a well-regulated and widely harmonized Air Traffic Management framework, supported by a worthwhile and viable Communications, Navigation and Surveillance (CNS) infrastructure.*

Here, ML is used for the recommendation portion of the project. Machine Learning helps in identifying the patterns in the data which might not be visible to the naked eye. It then is used to make useful predictions and correct recommendations based on the data into consideration.

Android provides a rich application framework that allows us to build innovative apps and games for mobile devices in a Java language environment. Here, Android is used to make a proper application interface of ATC.

Nomenclatures used: *Federation Aviation Administration (FAA), Air Traffic Control (ATC), Air Traffic Management (ATM), Machine Learning (ML), Ground Delay Program (GDP), Control Process (CP).*

INTRODUCTION

1. PROBLEM STATEMENT

1.1 Problem Definition

i. General: The problem is the structural mismatch between the nature of air traffic control and the way the federal government manages it.

ii. Uncertainty of Funding in a Political Environment: In recent years, funding uncertainties resulting from sequestration, government shutdowns.

iii. Lack of a Capital Budget: Lack of a stable funding stream makes planning for multi-year projects almost impossible.

As a result, we have seen significant delays and inefficiencies in modernization.

1.2 Problem Explanation

In the project, the major complication is traffic overpopulation we are dealing with (funding problem is also included but at a secondary level). There can be a bottleneck ranging from areas of air traffic to freight carriers. Apace with this, the problems faced in the ongoing framework, due to the constructions of roads, runways, etc. are added.

So, there is a lot of hazards involved in trying to maintain and look after the substantiality of the traffic and henceforth, controlling the air crowding. Thus, we have tried to make an application that will help passengers choose airlines intelligently and sensibly also the controllers to maintain stability in the system with as much ease as possible.

The aggrandized competence of the system to help control network traffic is additional. The METAR data is used to guide and monitor the real- time air traffic and just like that to give accurate results. This tracked information aids in the prognosis of the traffic denseness and then the network working is augmented. This renders collectible information (facts and news) for controlling traffic outflow, prediction of congestion and contracting the total of accidents in that network.

The advancement and expansion of know-how administration have influenced a large number of areas. Nonetheless, there are two opportunities to the scientific association on how to lead with an amount of data so big in real-time and bring about convenient results; and with Big Data available on how to meliorate the real-time decision support systems adopting historical information. There has been an escalation in unregulated data off late. To find which of them is fruitful and insightful is the task of the government.

2. METHODOLOGY

Recommendation systems have been emerging in contemporary years. During the time people used to accomplish it with electronic commerce shopping portals, now it can be practiced to any one thing ranging from webpages to blogs, search engines, and even websites. The two preeminent accession to physique these recommendation systems are:

1. **Collaborative Filtering:** CF perpetuates a database of ratings given by infrequent end-users for disparate products or flights (in this case). It then looks out for ratings that are analogous to alternative ratings and clubs them well-balanced and well-organized. Collaborative Filtering is very extensively used by many companies and to frame such a system, one needs to use not only the user data but also the product data.

2. **Content Based:** Content-based systems, as the name suggests, uses data only from the products or items incorporated. CB looks into the countenance of the components and then upholds an item with analogous features to the one which is being gone through.

Although, until now there is no authentic explanation of Big data, it can be elucidated formally for knowledge discovery in so big data structures. The ramification of analyzing big data is based on three factors: capability i.e., volume, velocity, and versatility so that they can make a base for business specialists to continue the decision process.

- **Volume:** The size of the data.
- **Velocity:** The change of speed from the oldest data to the newest one.
- **Versatility:** The total measure of sources of data and their understanding and merging.

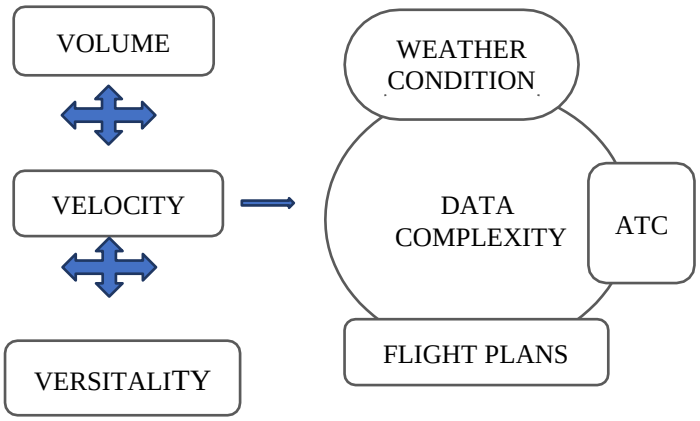

Figure 1: ATC and Big Data

2.1 HOW WE DO IT:

Our scenario of doing work is first to filtering and cleaning of data through Big data Hadoop. In technical terms, perfect sorting of data will take place at this very initial stage. Then, via Machine Learning, all the filtered data (or we can say datasets) will be sorted in a manner so as to apply algorithms and sub-methodologies of ML in order to achieve goals like smart learning, classifications, visualization and recommendations. At final stage, android will come in use to get the proper user-interactive interface to make app easy to use for third-party users or non-technical individuals with integrity, high-throughput and highest degree of efficiency. All the work will be done in a sequential order, if one step is missing then it will be very difficult to be on the same track again.

3. LITERATURE REVIEW

The ATM environment can be classified into 3 different sectors:

a) **Air Space Management:** ASM directs on broadening the proficiency of aircraft in the airspace, to provide adequate services for demand within the available structure.

b) **Air Traffic Control:** ATC focuses on controlling the aircraft flight, providing essential information that preserves the clause of safety.

c) **Air Traffic Flow Management:** ATFM centers on yielding information to uphold the air traffic percolate with safety and a shortened impact on the future approach of performing any task.

Recommendation: Predicts whether the airline reviewers recommend the airline to others that would be based on their experience. Based on airline reviews, a dataset gathered from various airlines, providing propaganda about the quality of an airline based on various factors like seat comfort, cabin crew, in-flight entertainment, ground crew, etc.

4. TOOLS AND TECHNOLOGIES

4.1 MACHINE LEARNING

Machine learning is the mechanism of data scrutiny that brutalizes penetrating exemplary buildings. It is a subsidiary of artificial intelligence positioned on the perception that systems can pick- up and grasp from data, identify patterns and generate decisions with nominal human intercession.

We evaluate the performance measure of various machine learning techniques in predicting the recommendation decision. Towards that goal, we perform supervised learning and learn from illustrated examples i.e., the digital reviews and the indication of the reviewer expressing the airline recommendation. In this context, we majorly concentrate on Naïve Bayes (NB) as a rather simple learning algorithm as well as Neural Network (NN) and Support Vector Machine (SVM) representing more space and time- consuming learning algorithms.

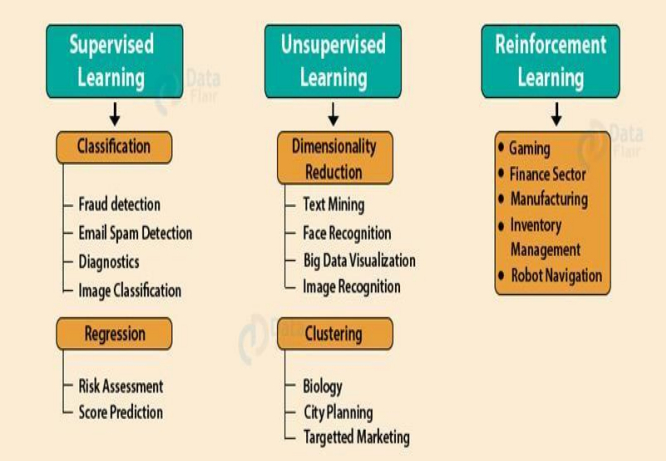

Figure 2: ML Algorithm Flowchart

Algorithms we are using in the project,

A. Naive Bayes Algorithm

Naive Bayes Classifier is a Supervised machine-learning algorithm that uses the Bayes' Theorem, which assumes that features are statistically independent. The theorem is based on the naive expectation that input valuables are self-determining of each other,

i.e. there is no means or measure to know anything about other values when given an additional variable.

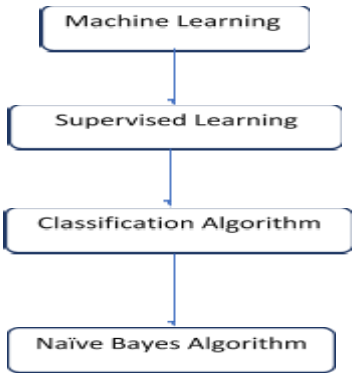

Figure 3: Naiive Bayes in ML

Bayes' Theorem is stationed on the tentative probability or in an uncomplicated manner, the prospect that an event (A) will occur given that another event (B) has already occurred.

Let's explain what each of these terms means.

- "P" is the symbol to denote the probability,
- P (A | B) = The probability of event A

  (interpretation) occurring given that B (signify) has come to pass

- P (B | A) = The probability of event B (signify) occurring given that A (interpretation) has appeared P(A) = The probability of event B (interpretation) occurring P(B) = The probability of event A (signify) occurring

B. Decision Tree Algorithm

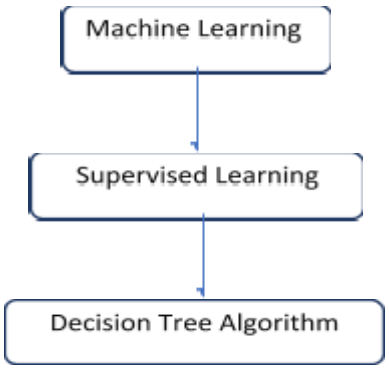

Figure 4: Decision Tree in ML

The two entities of tree are:

i. **Regression**: The cost function that is minimized to choose split points is the sum squared error across all training samples that fall within the rectangle.

ii. **Classification**: The cost function is used which provides an indication of how pure the nodes are, where node purity refers to how mixed the training data assigned to each node

C. Random Forest Algorithm

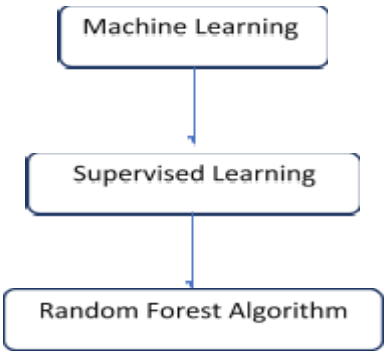

Figure 5: Random Forest in ML

Random forest is a supervised learning algorithm that comes in use for both coordination as well as retrogression.

As we know that a forest is made up of trees and more trees mean more robust forests. Similarly, random forest algorithm invents decision trees on data savors and then gets the foretelling from each of them and convincingly selects the most credible solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

4.2 HADOOP

Hadoop is an open-source framework designed for distributed storage and processing of very large data sets across clusters of computers. Hadoop consists of components including:

- Hadoop Distributed File System (HDFS), the bottom layer component for storage. HDFS breaks up files into chunks and distributes them across the nodes of the cluster.
- Yarn for job scheduling and cluster resource management.
- MapReduce for parallel processing.
- Commonlibraries needed by the other Hadoop subsystems.

Hadoop follows a Master-Slave architecture for the transformation and analysis of large datasets using the Hadoop MapReduce paradigm.

A	B	C	D	E	F	G	H	I	J
	Year	Month	DayofMor	DayOfWe	DepTime	CRSDepTi	ArrTime	CRSArrTin	UniqueCa
0	2008	1	3	4	2003	1955	2211	2225	WN
1	2008	1	3	4	754	735	1002	1000	WN
2	2008	1	3	4	628	620	804	750	WN
4	2008	1	3	4	1829	1755	1959	1925	WN
5	2008	1	3	4	1940	1915	2121	2110	WN
6	2008	1	3	4	1937	1830	2037	1940	WN
10	2008	1	3	4	706	700	916	915	WN
11	2008	1	3	4	1644	1510	1845	1725	WN
15	2008	1	3	4	1029	1020	1021	1010	WN
16	2008	1	3	4	1452	1425	1640	1625	WN
17	2008	1	3	4	754	745	940	955	WN
18	2008	1	3	4	1323	1255	1526	1510	WN
19	2008	1	3	4	1416	1325	1512	1435	WN
21	2008	1	3	4	1657	1625	1754	1735	WN

Figure 6: Snippet from a Dataset used

4.3 HADOOP TOOLS USED IN THE PROJECT

I. HIVE - Hive is a data warehouse system for data summarization and analysis and for querying of large data systems in the open-source Hadoop platform. It converts SQL-like queries into MapReduce jobs for easy execution and processing of extremely large volumes of data.

The three necessary purposes for which Hive is disposed of are data definition and depiction, data analysis, and data query. The query language, solely sustained by Hive, is HiveQL. This language translates SQL-like queries into MapReduce jobs for deploying them on Hadoop.

Applications of Hive: Hive is mainly used for data querying, analysis, and summarization. It helps improve developers' productivity which usually comes at the cost of increasing latency. Hive is an alternative to SQL and a sublime one indeed. It stands tall when compared to SQL systems implemented in databases. Hive has many user-defined functions that offer effective ways of solving problems.

It is conveniently viable to connect Hive queries to numerous Hadoop packages like RHive, RHipe, and even Apache Mahout. Also, it greatly helps the developer community work with complex analytical processing and challenging data formats.

II. PIG - Pig is a platform utilized to analyze large datasets consisting of high-level language for expressing data analysis programs along with the infrastructure for assessing these programs. Pig programs can be highly parallelized due to the virtue of which they can handle large data sets. Pig comes with a rich set of data types and operators to do different data operations. If the programmers hope to run a task in Pig then they need to use this language to write a Pig script. They then can use it to execute through various execution mechanisms like UDFs, Embedded, Grunt Shell. To produce the intended output these scripts will go through a series of transformations applied by Pig to give the desired output. Pig internally converts these scripts into a series of MapReduce jobs and therefore the programmers work is made much easier.

4.4 ANDROID

Android is used for Mobile app development. **Mobile apps development** means the core development of software particularly for the smart phones and other gadgets. Most popular and well-known mobile operating systems are iOS, android, windows and blackberry. Each operating system follows own rules and development procedures when to develop a particular application on different platform. So, it's very important for the developers to grip all the techniques

that are suitable for each platform when to develop an application.

As an example we can say that an android application can't run in windows or iOS platform. At present, chatting, cooking, matrimony, shopping, money making, share market news, banking- all at our hold.

So, the demands of the developers are highly needed. From this we get the idea of making an application through which one can book or check the status of both domestic and international flights of different airlines with respect to its timing, price (including taxes) and corresponding airport.

5. RESULTS AND CONCLUSIONS

This paper underlines data analysis on airline data sets. The paper addresses the usage of modern analytical tools. Hive on Big Data set which focuses on general services and requirements of any airport. Some of the instances are marked and highlighted with the images. The paper also provides an introduction to Map Reduce techniques that are internally taken care of by the underlying tools of the Hadoop System.

Here, we have focused on the handling and processing of the big data sets using the hive and pig components of the Hadoop ecosystem in a distributed environment. This work will further benefit the developers and business analysts in accessing and processing their user queries.

We have tried to implement several different machine learning algorithms - several of them are classification algorithms. These include Random forest, naiive Bayes, decision tree and a clustering algorithm k-means clustering too.

We have also made the android designing of our aimed final product with the input boxes for the departure and arrival airports. We are in the early stages of the process of getting the outputs from the machine learning algorithms to be linked to a dynamically, and fully functioning android application. This when happens, will produce results of airports not just based on price but also on the basis of the airline delays in the particular route and preference of the airlines by people.

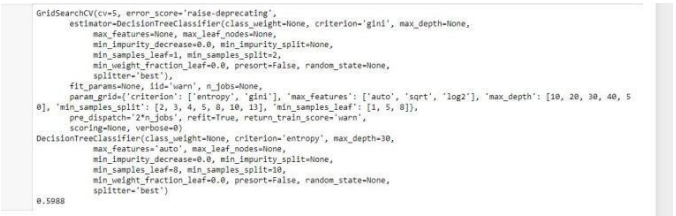

Figure 7: Results: Naiive Bayes

Figure 8: Results: Decision Tree Classifier

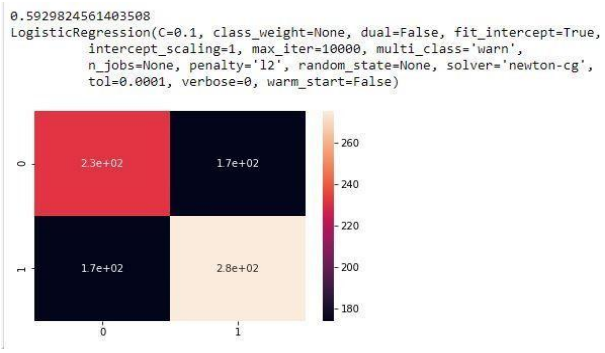

Figure 9: Results: Logistic Regression

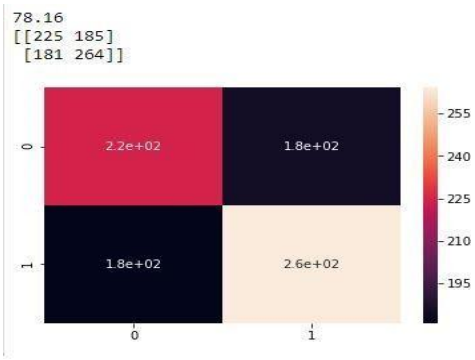

Figure 10: Results: K-means Classifier

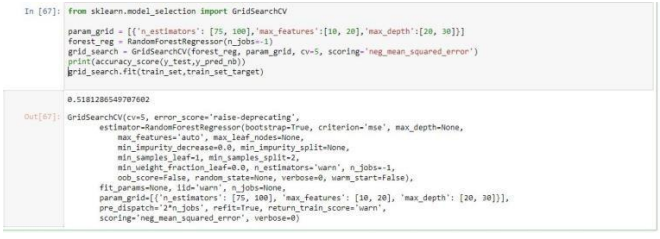

Figure 11: Results: Random Forest

The above are some of the results from the various Machine Learning algorithms that were implemented. Below shown, are some of the results for some US domestic airports.

As an illustration, we have taken two airports- DEN (Denver) and LAX (Los Angeles) airports. The results are shown as the following.

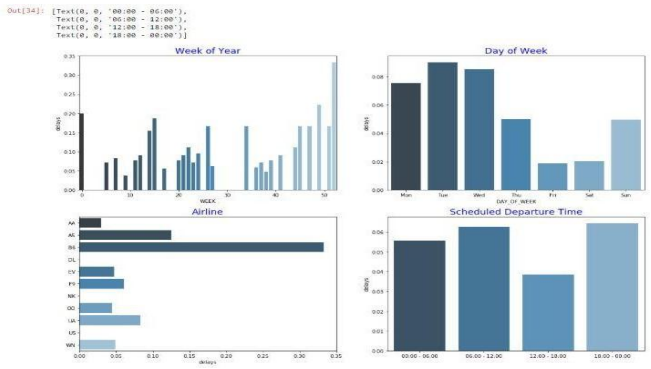

Figure 12: Results: DEN airport

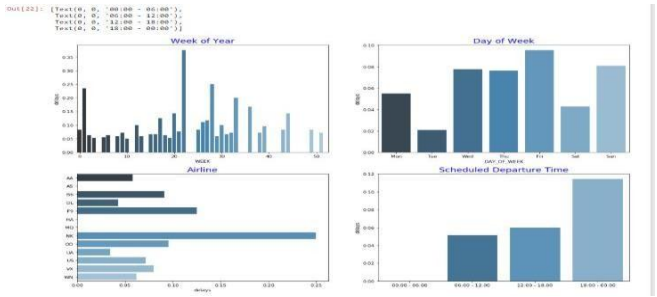

Figure 13: Results: LAX airport

Now shown are some of the designs made for the final application we have in our minds. When connected to the outputs we are getting from the machine learning algorithms, we hope to produce dynamic results for flight searches made on the android application.

In this research paper however, we stick to the outputs shown on the jupyter notebook and the made designing of our application leaving out the dynamic linking.

Figure 14: Results: Login page of App

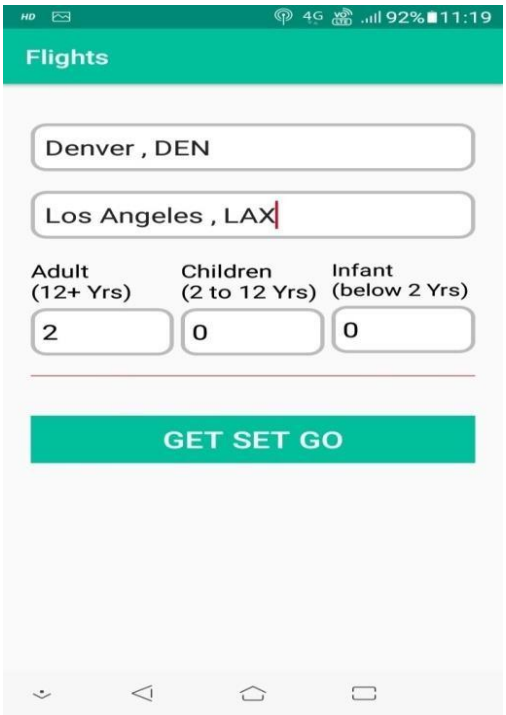

Figure 15: Results: Departure and Arrival Selection

6. REFERENCES

1. M. Abdel-Aty, C. Lee, Y. Bai, X. Li, and
2. M. Michalak. "Detecting periodic patterns of arrival delay", Journal of Air Transport Management, 13(6):355–361, Nov. 2007.
3. K. F. Abdelghany, S. S. Shah, S. Raina, and A. F. Abdelghany, "A model for projecting flight delays during irregular operation conditions", Journal of Air Transport Management, 10(6):385–394, Nov. 2004.
4. S. Ahmad Beygi, A. Cohn, Y. Guan, and P. Belobaba, "Analysis of the potential for delay propagation in passenger airline networks", Journal of Air Transport Management,14(5):221–236, Sept. 2008.
5. F.F. Reichheld, "The one number you need to grow", Harvard Business Review 81 (12)(2003) 46–55.
6. B.A. Sparks, H.E. Perkins, R. Buckley, "Online travel reviews as persuasive communication: the effects of content type, source, and certification logos on consumer behavior", Tourism Management 39(2013) 1–9.
7. Q. Ye, R. Law, B. Gu, W. Chen, "The influence of user-generated content on travel behavior: an empirical investigation on the effects of e-word-of-mouth to hotel online bookings", Computers in Human Behavior (2011) 634–639.
8. H.T. Rhee, S.-B. Yang, "How does hotel attribute importance varies among different travelers? An exploratory case study based on a conjoint analysis.