A PROJECT

ON

“Predection Of flight Cancellation Analysis”

SUBMITTED IN
PARTIAL FULFILLMENT OF THE REQUIREMENT
FOR THE COURSE OF
DIPLOMA IN
BIG DATA ANALYTICS FROM CDAC

SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY
‘Hinjewadi Phase-2 pune’.
MH-INDIA

SUBMITTED BY
Vaibhav Tejrao Wanare

UNDER THE GUIDANCE OF
Mr. Girish Gaikwad
Faculty Member
Sunbeam Institute of Information Technology, PUNE.

CERTIFICATE

This is to certify that the project work under the title ‘Prediction of Flight Cancellation Analysis’ is done by Vaibhav Tejrao Wanare in partial fulfillment of the requirement for award of Diploma in Big Data Analytics Course.

Mr.Girish Gaikwad
Project Guide

ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of Gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mr. Nilesh Ghule and Project Guide Mr.Girish Gaikwad. We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Vaibhav Wanare

TABLE OF CONTENTS

1. Introduction

When planning travels, especially in the phase of booking flights, it would be good for a customer to know the likelihood of his/her flight being cancelled. This useful information will help customers better plan their trips. To this end, the objectives of this project are two folds: analyze airline flight cancellation patterns and develop prediction models for flight cancellation. In this post, exploratory data analysis will be conducted to summarize and detect patterns of flight cancellation by airlines using the U.S. domestic flights data from 2008. In the next step, I will focus on the development of prediction models with respect to the likelihood of a flight being cancelled. Ensemble learning methods (e.g. random forests and boosting) will be used.

1.2 Problems Statements:

Predicting flight status is very vast and difficult to understand. It is considered too uncertain predictable due to huge fluctuation of the parameters regarding the flights and the airlines. Flight status prediction task is interesting as well as divides researchers and academics into two groups, those who believe that we can devise mechanisms to predict the flight status and those who believe that the flights status is efficient and whenever new information comes up the absorbs flight status by correcting itself, thus there is no space for prediction.
Planning for a trip but at a bad time can have disastrous result, while booking a flight at the right time can bear good vibes. Business class of today are facing this problem of flights cancellation as they do not properly understand as to which flight to reserve or which flights should avoid. So, the purpose project will reduce the problem with suitable accuracy faced in such real time scenario.

1.2 Relevant current/open problems:

1. Data-are-humongous, nowadays we are seeing a rapid-explosion of flight cancellation. They arise from all different-sources.

2. Predictive modeling is important since the basic op management process going from the vendor of the airlines system to the customers hands, takes some time. Most passengers cannot wait for to elevate and then give a reaction. Instead, they make-up their mind and plan according to future consideration so-that they can react accordingly.

3. Generally predictive -lead to good-ops-and great-levels of customer satisfaction, while bad prediction will definitely-lead to costly ops and worst-levels of customer satisfaction.

4. A confusion for the predictive modeling is the horizon, which is, how distant in the future will the predict project? As a simple rule, the away into the future we see, the more blurry our vision will become – distant predictions will be inaccurate that short-range forecasts.

1.3 Technical analysis:

In contrast to fundamental analysis, technical analysis does not try to gain deep insight into a airlines systems. It assumes the available public information does not offer a competitive predictive advantage. Instead, it focuses on studying a airlines data released by and on identifying patterns in the chart. The intention is to recognize trends in advance and to capitalize on them.

1.4 Goal:

The goal was to build a system capable of the following tasks:

1. Collecting fundamental and technical dataset
The system should consist of a dataset related to title. Furthermore, it should be able to collect the pattern from the dataset which will help to design suitable ML model.

2. Simulating strategies of models
The system/model should offer ways to specify and simulate fundamental and technical prediction strategies. Additionally, combining two or more model strategies should be possible.

3. Evaluating and visualizing results
The system should evaluate and visualize the parameter of the performance and attribute considered. This allows a comparison to be made between technical, fundamental and the combined approaches.

1.5 Project Goals and Scope:

The current airline environment is very competitive and dynamic. Maintaining consistent profitability requires that appropriate trade offs be made between the often competing objectives within planning, marketing and operations. A chief goal of this project is to add to the academic understanding of a flight cancellation prediction. The hope is that with a greater understanding of how and what factors affect the cancellation of the flight. The project will evaluate some existing strategies from a rigorous scientific perspective and provide a quantitative evaluation of strategies. It is important here to define the scope of the project. The project deals with the prediction on the data of year 2008. However an attempt can made to make this project with

recent data and considering the proper attributes and the features. This can certainly enable the prediction to be perfect and more durable. This project will focus exclusively on predicting status of the flight (will be cancelled or not) and the

probability of cancellation and analysis of previous record of cancellation. More so, the project will analyze the accuracies of these predictions.

1.6 Overview of proposed solution approach:

1. Basically the main objective of this project is to collect the airlines information for some previous years and then accordingly predict the results (predicting what would happen next). So for we are going to use of two well-known techniques Machine Learning and data mining for flight cancellation prediction. Extract useful information from a huge amount of data set and data mining is also able to predict future trends and behaviors.

2. As far as the solutions for the above problems, the answer depends on which way the forecast is used for. So the procedures that we will be using have proven to be very applicable to the task of forecasting product demand in a logistics system/ Binary classification systems. Many techniques, which can prove useful for forecasting-problems, have shown to be inadequate to the task of demand forecasting in logistics systems(Binary systems)network. Therefore, combining both these techniques could make the prediction more suitable and much more reliable.

3. As far as the solutions for the above problems, the answer depends on which way the forecast is used for. So the procedures that we will be using have proven to be very applicable to the task of forecasting product demand in a logistics system. Many techniques, which can prove useful for forecasting-problems, have shown to be inadequate to the task of demand forecasting in logistics systems.

1.7 Novelty/Benefits:

Flight delays are an important subject in the literature due to their economic and environmental impacts. They may increase costs to customers and operational costs to airlines. Apart from outcomes directly related to passengers, flight cancellation prediction is crucial during the decision-making process for every player in the air transportation system. Also the knowing status of the flight enables a businessman/entrepreneur to accordingly plan the meeting.

2. Product Overview and Summary

2.1 Purpose:

The aims of this project are as follows:

1. To identify factors affecting flights cancellation
2. To generate the pattern from large set of data of airlines for prediction of flight cancellation.
3. To predict weather the flight is going to cancel or not based on the attribute selected from the dataset.
4. To provide analysis for users.

The project will be useful for player/businessman to check out the status based on the various factors. The project target is to create web application that analyses previous airlines data and implement these values in data mining/Machine Learning algorithm to determine status of the flights in near future with suitable accuracy. These predicted and analyzed data can be observed by individual to know the flight status. Company and industry can use it to breakdown their limitation and enhance their ability of serving the customer . It can be very useful to even researchers, market makers, government and general people.

2.2 Scope:

Flight can be cancelled by different sources and affect airports, airlines, en route airspace or an ensemble of them. For analysis purposes, one may assume a simplified system where only one of these actors or any combination of them is considered. Some work focused on airports to predict status for all departs considered all airlines and en route airspace indifferently. Airports are also the focus when the objective is to investigate their efficiency based on cancellation of all carriers. On the other hand, only airlines are considered when comparing the performance of two airlines under the same conditions . An ensemble of airport and en route airspace were studied to understand the relationship between congestion and cancellation. Others considered airports and airlines as well to evaluate capacity problems and airlines decisions. There are many possibilities to ensemble scopes. This becomes important when studying the dynamics of air transportation systems, mainly when targeting root cancellation.

3. Requirements and Feasibility study

3.1. Feasibility Study:

Simply, flight status cannot be accurately predicted. The future, like any complex problem, has far too many variables to be predicted. The flights status is a place where relationship between customers and airlines business associates converge.Larger the issues of the airlines lesser is the customer satisfaction. When there are more issues of any airlines then it is ot assured as quality assured and is probably avoided.. So, there is a factor which ask people to try for some other option. It has more to do with emotion than logic. Because emotion is unpredictable, flight cancellation will be unpredictable. It's futile to try to predict weather flight will be cancelled or not. They are designed to be unpredictable.There are some fundamental airlines indicators by which a flight status can be estimated. Some of the indicators and factors are: day of week, day of month, arrival time, departure time, CRS arrival time, CRS departure time, elapsed time, CRS elapsed time, weather delay, air time etc. Some of the parameters are available and accessible on the web but all of them aren't. So we are confined to use the variables that are available to us.The proposed system will not always produce accurate results since it does not account for the human behaviors. Factors like carrier number, unique carrier, taxi-in,taxi-out, flight number cannot be taken into account for relating it to the change in prediction of the status. Also the variable like weather delay , NAS delay cannot be taken into consideration as they contain more no of the bad record.The objective of the system is to give a approximate idea whether flight will be cancelled or not. It does not give a long term forecasting of a status. There are way too many reasons to acknowledge for the long term output of a current status.Many things and parameters may affect it on the way due to which long term forecasting is just not feasible.

3.2. Requirement Analysis:

After the extensive analysis of the problems in the system, we are familiarized with the requirement that the current system needs. The requirement that the system needs is categorized into the functional and non-functional requirements.
These requirements are listed below:

1. Functional Requirements
2. Non-Functional Requirements

3.2.1 Functional Requirements:

Functional requirement are the functions or features that must be included in any system to satisfy the business needs and be acceptable to the users. Based on this, the functional requirements that the system must require are as follows:

1. The system should be able to generate a approximate flights status(cancelled/not).

2.The system should create accurate model from the dataset provide in consistent manner. :
3. Prior to application of flight cancellation estimator, the data is updated by the latest values for the cross validation.
4. The user can look previous data Information which was collected.

3.2.2 Non-Functional Requirements:

Non-functional requirement is a description of features, characteristics and attribute of the system as well as any constraints that may limit the boundaries of the proposed system. The non-functional requirements are essentially based on the performance, information, economy, control and security efficiency and services. Based on these the non-functional requirements are as follows:

1. The system should provide better accuracy.
2. The system should have simple interface for users to use.
3. To perform efficiently in short amount of time.

1. Reliability:
The reliability of the product will be dependent on the accuracy of the dataset of airlines, how much flights were cancelled. Also the data used in the training would determine the reliability of the software.

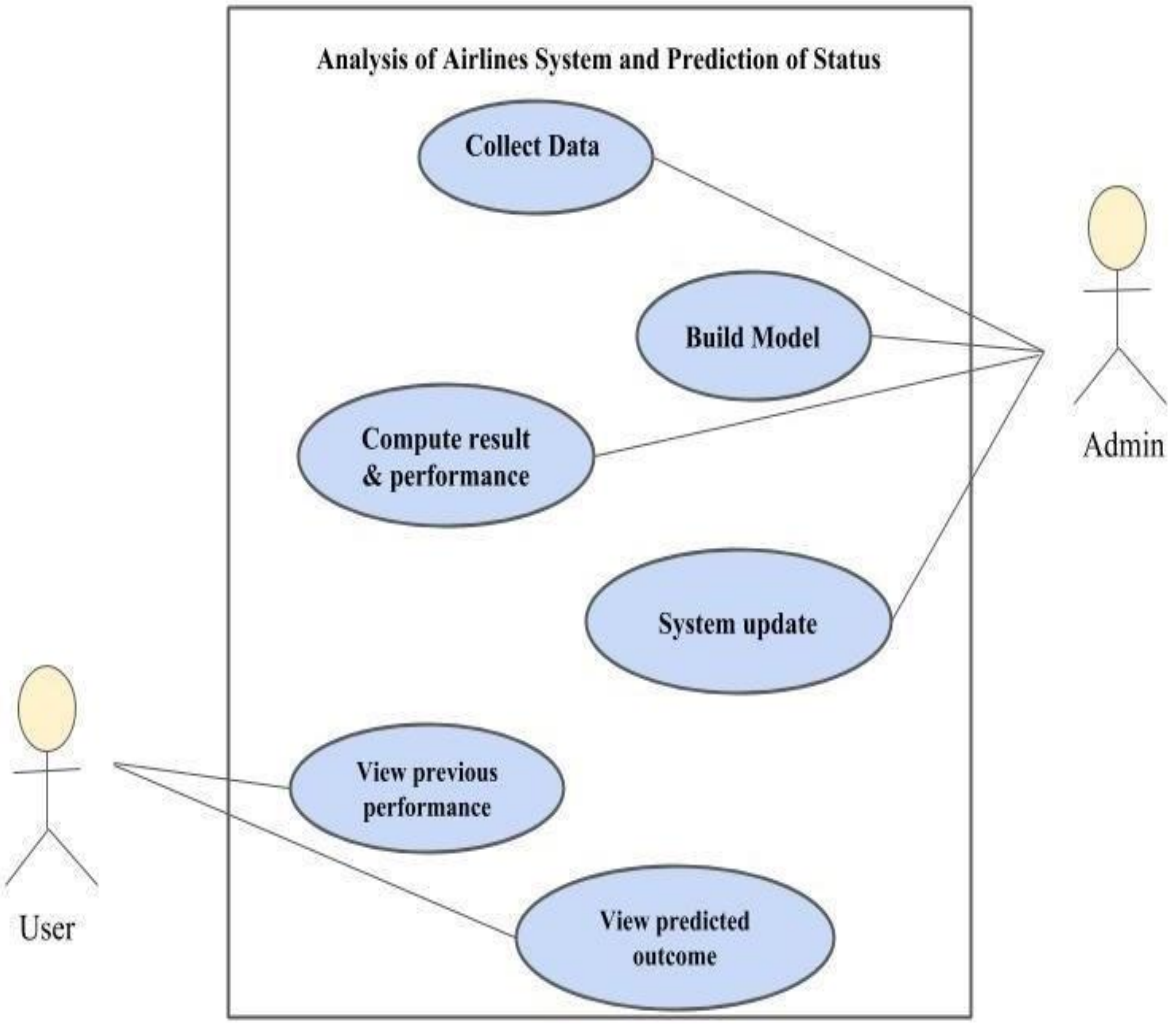
2. Security:
The user will only be able to access the website using fields details and will not be able to access the computations happening at the back end.

3. Maintainability:
The maintenance of the product would require training of the software by recent data so that there commendations are up to date. The database has to be updated with recent values.

4. Portability:
The website is completely portable and the recommendations/prediction completely trustworthy as the data is dynamically tested.

4. System Design and Architecture

4.1 Use Case Diagram:



Use Case Id	Use case name	Primary	Scope	Complexity	Priority
1	Collect Data	Admin	in	High	1
2	Build Model	Admin	in	High	1
3	Compute result & performance	Admin	in	High	1
4	System update	Admin	in	High	1
5	View previous performance	User	in	Medium	2
6	View predicted outcome	User	in	High	1

Use case ID: 1
Use case name: Collect data
Description: Every required data will be available in csv file. System will be able to collect the data for model.

Use case ID: 2
Use case name: Build model
Description: Every required data will be available in csv file. System will be able to collect the data for model. Based on data suitable ML model is build.

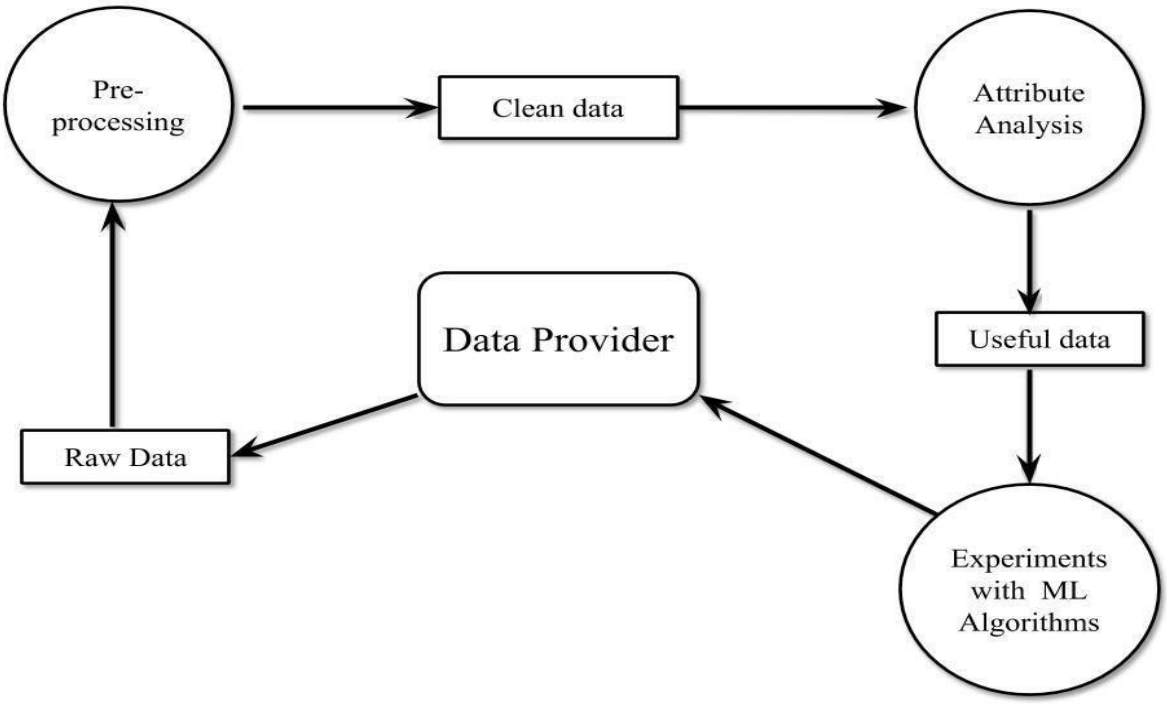
Use case ID: 3
Use case name: Compute result and performance
Description: Prediction result will be handled and generated by System. The system will be built, through which the result of prediction and system performance will be analyzed.

Use case ID: 4
Use case name: System update
Description: With the change of airlines regular update of system is required. Based on this there we predict the status of flight.

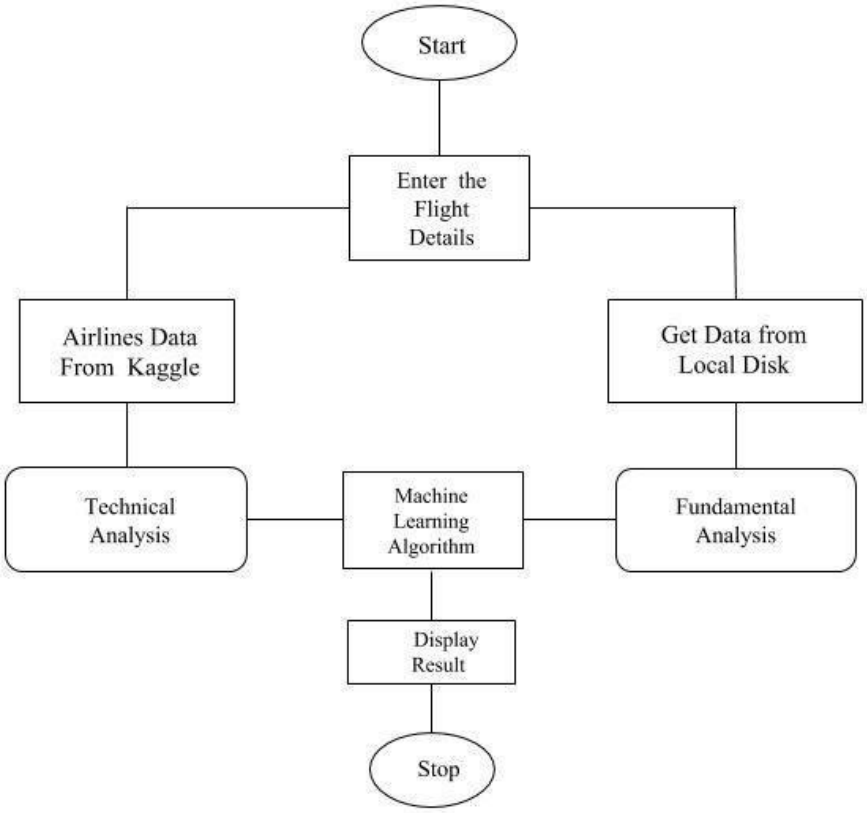
Use case ID: 5
Use case name: View previous performance. Description: Previous record of airlines among the source and destination can be viewed by user.

Use Case ID: 6
Use Case Name: View predicted outcome
Description: This use case is important in whole project. The key feature of this project is to predict the status of flight. Thus, this will be available in user interface and viewer can observe them.

4.2 System Flow Diagram:



4.3 Control Flow Diagram :



4.4 Proposed Algorithm:

Logistic Regression

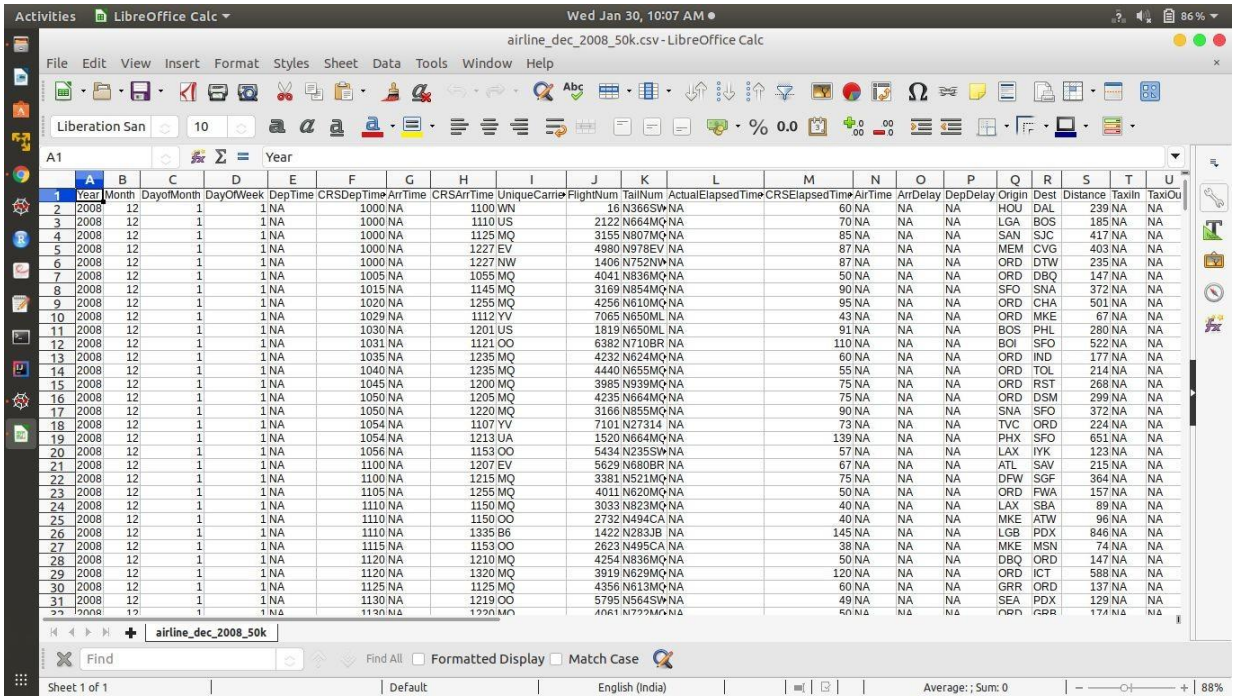
Logistic regression belongs to the group of regression methods for describing the relationship between explanatory variables and a discrete response variable. A logistic regression is proper to use when the dependent is categorized and can be applied to test association between a dependent variable and the related potential factors, to rank the relative importance of independents, and to assess interaction effects (Allison, 1999). Binary logistic regression is used when the dependent variable Y can only take on two values (such as low delay vs high delay).

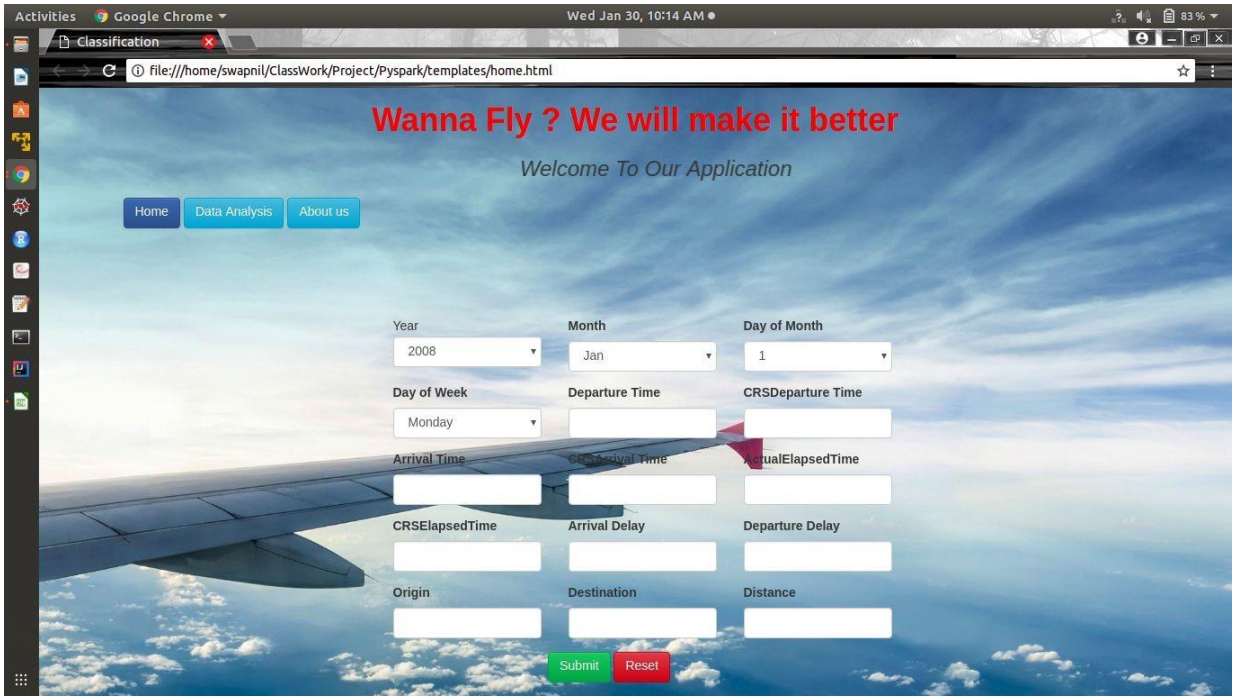
Random Forest Classifier :

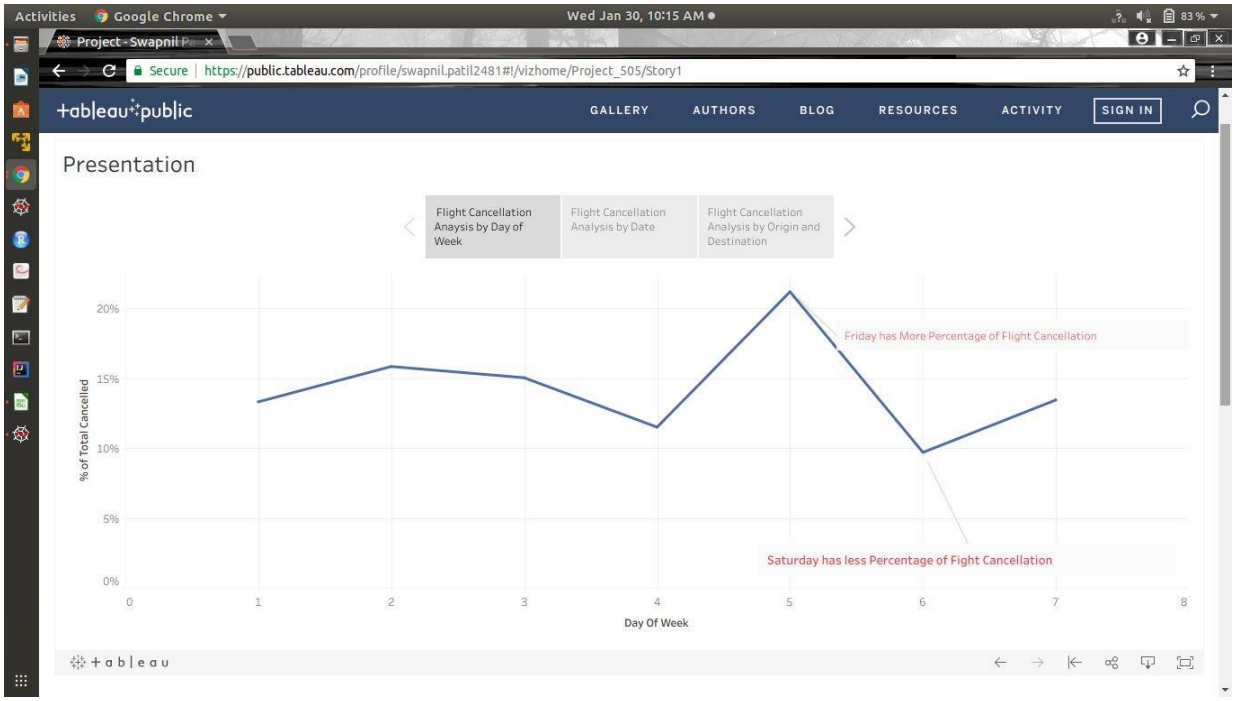
Random forest Decision tree learning is one of the most popular techniques for classification. Its classification accuracy is comparable with other classification methods, and it is very efficient. ID3 presented by Quinlan (1986), C4.5 presented by Quinlan (1993) and CART presented by Breiman et al (1984) are decision tree learning algorithms. Details can be found in article of Han et al (2006). Random forest belongs to the category of ensemble learning algorithms. It uses decision tree as the base learner of the ensemble. The idea of ensemble learning is that a single classifier is not sufficient for determining class of test data. Reason being, based on sample data, classifier is not able to distinguish between noise and pattern. So it performs sampling with replacement such that given n trees to be learnt are based on these data set samples. Also in our experiments, each tree is learnt using 3 features selected randomly. After creation of n trees, when testing data is used, the decision which majority of trees comes up with is considered as the final output. This also avoids problem of over-fitting.

5. FINDINGS AND CONCLUSIONS

The system evaluation on the flight status from airlines system is carried out. For given days of week, days of month, day's arrival time, CRS arrival time, CRS elapsed time, Actual Elapsed time, Departure delay, CRS departure delay, Air Time, and adjacent values along with the Origin and Destination, our forecaster will forecast if flight will be cancelled or not.Our predictive model is evaluated on airlines data on the historical data over the period of December 2008.The accuracy of the system is measured as the percentage of the predictions that were correctly determined by the system. For instance, if the system forecasts an 0 value and the flight is actually not cancelled , it is supposed to be correct, otherwise, if the value is 1, it is assumed to be wrong.Following Airlines dataset is taken as sample training data over the period of 30 days. Corresponding rates file is also provided along with this. Predictions using related attribute is also shown .




Activities
Google Chrome
Wed Jan 30, 10:14 AM
83 %
Classification
file:///home/swapnil/ClassWork/Project/Pyspark/templates/home.html
Wanna Fly ? We will make it better
Welcome To Our Application
Home
Data Analysis
About us
Year
2008
Month
Jan
Day of Month
1
Day of Week
Monday
Departure Time
CRSDeparture Time
Arrival Time
CRSArrival Time
ActualElapsedTime
CRSElapsedTime
Arrival Delay
Departure Delay
Origin
Destination
Distance
Submit
Reset



5.1 CONCLUSION

Evaluating the Flights status has at all times been tough work for analysts. Thus, we attempt to make use of vast written data to forecast the status of flight in dices. If we use techniques of text and numeric data analysis the accuracy in predictions can be achieved. Business class people, Sports player or even a normal customer can use this prediction model to take t decision by observing prediction of the model.

5.2 FUTURE WORK

1. A system/ model can be extended so that it could predict the best flight among the two route. That is we can build the flight recommendation can be build.

2.If we can collect the recent data, accuracy can be increased or we can use a technique of web scraping for the purpose of obtaining the training/testing dataset.